**Discrimination of phytochrome dependent light inducible from non-light inducible plant genes.
Prediction of a common light-responsive element (LRE) in phytochrome dependent light inducible
plant genes**

Ulrich Grob* and Kurt Stüber¹

Max-Planck-Institut für Immunbiologie, Stübeweg 51, D-7800 Freiburg i.Br. and ¹Max-Planck-Institut
für Züchtungsforschung, Egelspfad, D-5000 Köln 30, FRG

Received September 17, 1987; Revised and Accepted November 10, 1987

ABSTRACT

We aligned 14 5'-leading sequences of small subunit ribulose-1,5-bisphosphate carboxylase (rbcS) genes. A strong consensus sequence ("CCTTATCAT") was located directly upstream of the TATA-box. The occurrence of this motif in other light dependent phytochrome regulated plant genes led to the calculation of two consensus matrices. With these two matrices we are able to distinguish almost all known light induced plant genes which are phytochrome regulated from non-light induced plant genes indicating, that all these genes share a common light-responsive element (LRE). The results obtained by computer analysis are discussed with regard to experimental data.

INTRODUCTION

Phytochrome dependent light regulation is a major control mechanism in plants (1). Phytochrome is one of the three known light receptors in plants which have been shown to influence gene expression (2). The work presented here focusses mainly on genes controlled by this receptor. The transcriptional regulation of a number of nuclear genes encoding for small subunit of ribulose-1,5-bisphosphate carboxylase (rbcS) (3,4,5), chlorophyll a/b binding protein (3,4,5,6,7), protochlorophyllide reductase (7,8), chalcone synthase (9), rRNA (4,10,11) and several as yet unidentified gene products (4) has been shown to be under the control of the phytochrome system (5,7,12). Although the light stimulus is communicated via phytochrome, the molecular mechanism of the signal transfer remains unclear. The genes responsible for phytochrome production are negatively controlled by their own gene products (13,14).

In an attempt to elucidate structural elements common to

phytochrome regulated genes, we have devised a set of computer programs described elsewhere (17).

After locating a common sequence feature shared by all *rbcS* genes, we found this motif in addition in almost all phytochrome regulated genes.

These results are examined further in the light of experimental data.

HARDWARE AND SOFTWARE

The analysis was performed on a VAX 11/750 computer from Digital Equipment Corporation with the operating system VAX/VMS. Programs used were of the Genetics Computer Group (University of Wisconsin) (15) and our own program package GENEXPERT (16,17).

Sequence data were taken from the EMBL Nucleotide Sequence Data Library (European Molecular Biology Laboratory, Postfach 10.2209, D-6900 Heidelberg, West Germany), the GenBank™ Genetic Sequence Data Bank (Los Alamos National Laboratory, T-10, Mail Stop 465, Los Alamos, NM 87545, USA) and from our own plant gene library.

METHODS AND RESULTS

We attempted to find a common sequence feature shared by all *rbcS* genes, because from this gene family already about 30 different sequences are known.

Sequence Alignment

The program LINEUP (15) was used to align the 5'-upstream regions of those 14 *ssu-RubisCO* genes for which sufficient sequence data from the putative regulation regions are available. This alignment contains several blocks of homology (data not shown). The longest block is located around the TATA-box. This block and the corresponding consensus sequence derived by LINEUP is shown in figure 1.

Because LINEUP only allows manual alignments, we then estimated the probability, that this alignment could occur by chance within a sequence of definite length. Following von Hippel (18), we modified formula 1 from Boucher and Bryan (19) so that we were able to estimate the probability (P) of the occurrence of an oligomer of the length (L) with (M) matches, (B) blanks or gaps

sequence name:	sequence	reference:
gmrubps1	CGACACTATTATATATAGCAAGTTTGAGCAGA	PMB 7:451(1986)
lerbcs1	CCTTATCATTATATATAAGGGGACGATACCACTG	PNAS 83:3880(1986)
lerbcs3a	CCTTATCCTTATCATTATATATAGGGGAGAGACTAGAA	"
ntrubss	CCTTATCATTATATATAGGGTGGTGGGCAACT	NAR 13:2373(1985)
perbcs11	GCTTAACATTATATATAGAAGGGGGCACTATA	NAR 14:3325(1986)
ssu112	CTTTATCATTATAAATAGAGGGGGGAAGACTAC	EMBOJ 4:3055(1985)
ssu301	AAAAACCTTATCATTATATAAAGGGGTGTGGTGGGCAA	"
ssu491	CCTTATCACTATAAATAGAAAGTGGGACTAGGA	"
ssu611	CCTTTTCATTATATATAGGACAGGAGTCATAG	"
ssu911	CCTTATCACTATAAATAGATATTGCAAATGTC	"
strbcs1	CCTTATCATTTTCATTATATAAAGGGTAGTGGGCATCA	pers. comm. 1
strbcs2a	CCTTATCATTATATATAGAGGGGGAGACTACA	"
strbcs2b	..CCTTATCATCATTATATATACAAGGGGATACTAGA	"
strbcs2c	CCTTATCACTATATATACAGGGGGTTCCTAATA	"
consensus	ccttatcAttTATAtatagagggggg.acaa.a...	

Figure 1. Part of the alignment of 5'-upstream sequences of 14 rbcS genes. The major homology block around the TATA-box is shaded. The TATA-box sequences are boldfaced. The last line gives the consensus sequence derived by LINEUP (15). The 5-base shift or partial duplication of the upstream motif "CCTTATCAT" of the sequences LERBCS3A and STRBCS1 and the 3-base shift of the sequence STRBCS2B clearly can be seen. The abbreviations of the reference sources are explained in figure 4.

and (L-B-M) mismatches within a sequence of length (S). It is assumed that the investigated sequences are double stranded, not palindromic and that the A, C, G and T contents are not much different from 25% each. Terminating blanks or gaps are not taken into consideration. With these restrictions formula 1 of (19) is transformed to:

$$P = 1 - (1 - \left(\frac{L-B}{M}\right) * \left(\frac{L-2}{B}\right) * 0.25^M * 0.75^{(L-B-M)} * 1^B)^{(2 * (S-L+1))}$$

The consensus sequence derived by LINEUP was taken as a reference sequence in comparison to which the matches, gaps and mismatches of every aligned sequence were counted. The alignment probabilities range from 1,67% for GMRUBPS1 to 0,0000005% for PERBCS11A. These results show that it is highly unlikely that these alignments occur by chance, thereby suggesting that they may reveal conserved structures of biological relevance.

On the other hand, Boucher and Trifonov (20) have shown that the only well conserved consensus sequence in a representative set of 168 eukaryotic POL II promoters is the sequence "TATAAA"

*	Matrix Name: light1.imat													*

Pos:	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7
G	41	1	6	10	4	2	4	2	2	2	4	8	19	1
A	20	6	9	4	9	3	6	7	10	7	25	10	14	2
T	3	3	12	28	6	3	2	3	5	3	17	7	5	3
C	3	5	1	4	2	6	9	8	5	4	2	1	2	3
Total	67	16	29	46	21	13	21	19	22	15	48	26	39	9

Pos:	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
G	1	5	4	4	16	3	0	0	30	0	0	0	0	0
A	1	5	3	9	100	38	6	12	0	0	200	7	0	100
T	1	25	2	3	5	38	6	5	0	200	0	100	0	7
C	1	5	1	3	0	0	13	11	3	0	0	0	200	0
Total	4	42	10	19	121	80	150	133	85	200	200	176	200	176

Pos:	8	9	10	11	12	13	14	15	16	17	18	19	20	21
G	0	3	10	12	3	4	0	1	10	19	3	17	4	1
A	0	4	11	1	48	22	22	7	8	47	3	29	13	1
T	10	16	6	12	9	75	45	8	7	3	4	9	8	1
C	7	8	1	5	9	0	7	7	1	3	9	0	2	1
Total	176	31	27	32	69	101	73	24	26	72	19	55	28	3

Pos:	22	23	24	25	26	27	28							
G	49	19	8	4	21	2	1							
A	6	8	3	2	8	3	3							
T	12	36	7	3	3	7	2							
C	3	0	2	7	5	4	4							
Total	71	63	19	15	36	15	10							

Figure 2. The matrix LIGHT1.IMAT derived from 23 sites in 5'-3' forward direction. The core values corresponding to the consensus sequence are inversed. The algorithm used to calculate this information type matrix is described in (17).

(TATA-box), indicating that the homologous TATA-box-flanking sequences of our alignment are specific for the sequences used in this alignment. The regions 5' of the TATA-box are highly homolo-

*	Matrix Name: light1rev.imat													*

Pos:	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7
G	0	4	3	2	3	5	1	1	9	3	6	9	61	5
A	0	4	34	2	1	14	1	3	18	34	3	20	23	9
T	0	3	11	4	1	22	1	2	30	11	6	25	0	9
C	0	1	6	2	2	2	1	3	0	6	2	0	5	1
Total	0	12	54	11	7	42	4	9	57	54	17	53	90	25

Pos:	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
G	3	11	7	2	1	7	0	8	10	0	0	0	0	13
A	8	34	8	3	1	34	10	8	0	20	0	20	10	17
T	33	3	6	10	3	23	0	3	9	0	20	0	26	0
C	8	6	1	5	2	0	9	3	0	0	0	0	0	0
Total	52	54	22	21	7	63	171	52	171	200	200	200	126	150

Pos:	8	9	10	11	12	13	14	15	16	17	18	19	20	21
G	10	0	2	3	2	3	6	2	4	1	3	1	2	1
A	0	16	4	34	1	1	14	1	6	4	10	3	4	6
T	9	16	18	11	3	4	8	3	3	6	9	4	3	8
C	0	11	14	6	1	3	1	1	2	7	3	4	2	7
Total	171	43	39	54	7	11	29	7	15	19	24	12	10	22

Pos:	22	23	24	25	26	27	28							
G	1	1	2	11	2	1	3							
A	7	0	6	1	10	4	43							
T	6	1	6	8	20	3	7							
C	8	0	3	5	6	3	10							
Total	22	2	17	26	37	11	63							

Figure 3. The matrix LIGHT1REV.IMAT derived from 19 sites in 3'-5' reverse direction. The core values corresponding to the consensus sequence are inversed.

gous and show the consensus "CCTTATCAT". The sequences STRBCS1 and LERBCS3A show a 5-base shift or a partial duplication of this consensus sequence. In addition STRBCS2B shows a 3-base shift of

this sequence. The distance between the "CCTTATCAT" consensus and the TATA-box varies from 0 bp to 5 bp in the rbcS sequences.

Creating Consensus Matrices

In order to assess the frequency and distribution of the derived consensus, we investigated all plant sequences of the EMBL and GenBankTM libraries and additional plant sequences with the program FIND of the GCG program package (15), searching for the "CCTTATCAT" consensus sequence under allowance of one mismatch. From the sites found we have created two consensus matrices of the information type (17) using the program DYCOM from our GENEXPERT program package (16,17). In comparison to the results obtained with a single matrix, the results obtained by two different matrices proved to be much more informative, reflecting the fact that the matrices obtained for the two orientations of the 5'-CCTTATCAT-3' consensus are not identical, due to the limited nature of the database from which they were

sequence name:	seq. encodes for:	predicted light-responsive element	pos:	score:	f/r:	reference:
LIGHT INDUCED, POSITIVE REGULATED						
ftrbcs	rbcS	GAGTACCCCTCAGCATGGAT	435	44	f	NAR 15:1875(1987)
gmrbps1	rbcS	tcacctccacatggtaccta	449	33	r	PMB 7:451(1986)
gmrbps1	rbcS	GCATACCCCAATGGCTTCAT	1618	45	f	"
gmrbps4	rbcS	tgacacttttttagtaattt	112	33	r	"
gmrbps4	rbcS	tcacttccacatggtaccta	468	36	r	"
gmrbps4	rbcS	GCATACCCCAACGGCTTCAT	1644	44	f	"
harubps	rbcS	GAGTACCCCAAGGCTGGAT	463	45	f	NAR 15:1328(1987)
lecab3c	chloroA/B bind prot	aaagcagataatgatattct	62	38	r	Gene 40:247(1985)
lerbcs1	rbcS	XXXXXXXXXXCGCTAAAA	-11	41	f	PNAS 83:3880(1986)
lerbcs2a	rbcS	GCATACCCCAAGCATGGGT	462	37	f	"
lerbcs3a	rbcS	ATGGGAAGTTAAAGTGAA	38	38	f	"
lerbcs3a	rbcS	AAAGTTAAAGTGAAACCTTA	43	37	f	"
lerbcs3a	rbcS	GCTTACCCACAGGCATGGGT	766	43	f	"
nprbcs8b	rbcS	GATATGCGAAATATGTAA	979	43	f	MOG 205:193(1986)
nprbcs8b	rbcS	GCATACCCACAGGCTGGGT	1897	39	f	"
nsrubl	rbcS	GCATACCCACAGGCTGGAT	508	45	f	Bio 66:539(1984)
ntrbpc	rbcS	GCATACCCACAGGCTGGAT	166	45	f	FEBSL 154:145(1983)
ntrubcs	rbcS	GCATATAGGAAATATGTAA	925	43	f	NAR 13:2373(1985)
ntrubcs	rbcS	GCATACCCACAGGCTGGAT	1889	45	f	"
osrrn17s	17s rRNA	GTCCGGAGAAAGTCCATTGA	1725	38	f	NAR 12:5441(1984)
pecab22r	chloroA/B bind prot	gaatagatagtgattatca	90	37	r	NAR 13:2503(1985)
pecab91r	chloroA/B bind prot	AAACAAGATAATGATATTCA	81	38	r	"
pecab25	chloroA/B bind prot	gaacccgatagtgattatct	77	37	r	"
perbcs08	rbcS	ATATGTGAATATGTAA	370	41	f	NAR 14:3325(1986)
perbcs08	rbcS	GCTTAGTATTTACTAGCAAT	726	39	f	"
perbcs08	rbcS	GCTTACCCCAATGCATGGAT	1802	38	f	"
perbcs11	rbcS	GCTTACCCCAATGCCTGGAT	1093	39	f	"
pecab80	chloroA/B bind prot	ATGTGAATTATTATCATTA	2075	37	f	PNAS 81:2960(1984)
pach01	rbcS	atgtgtgaaggttaata	513	36	r	EMBOJ 4:2723(1985)
parbcs3a	rbcS	cactgtctcattgtgggtctt	208	37	r	Nat 287:692(1980)
parbcs3a	rbcS	aggggttagagaacaactt	711	33	r	EMBOJ 5:2063(1986)
consensus core sequence:		(forward)				
		(reverse)				

sequence	seq. encodes for:	predicted light-responsive element	pos:	score:	f/r:	reference:
name:						
LIGHT INDUCED, POSITIVE REGULATED						
parbcs3a	rbcS	GCTTACCCCCAAGCTTTGGT	1437	42	f	EMBOJ 5:2063(1986)
parbcs3a	rbcS	ggatggataaaggttaataa	1638	36	r	"
parbcs3c	rbcS	ttatgttcttcacacatga	104	35	r	"
parbcs3c	rbcS	GCTTACCCCTCAAGCTTTGGT	945	39	f	"
parbcs3c	rbcS	ggatgtataaaggttaataa	1140	37	r	"
parbcs36	rbcS	cattatcctattgttggtgca	771	38	r	Book 1
parbcs36	rbcS	GCTTACCCCTCAAGCTTTGGT	1612	39	f	"
parbcs80	rbcS	cattatcctagtgttggtgca	847	37	r	"
parbcs80	rbcS	GCTTACCCCTCAAGCTTTGGT	1690	42	f	"
parc01	rbcS	tattatcctattgttggtgca	849	38	r	EMBOJ 3:1671(1984)
parc01	rbcS	GCTTACCCCTCAAGCTTTGGT	1698	39	f	"
parc01	rbcS	ggatggagaaggttaataa	1903	36	r	"
parubp15	rbcS	GCTTACCCCTCAAGCTTTGGT	376	39	f	JBC 258:1399(1983)
pw9	rbcS	ctttctccatgatccggag	529	33	r	Btc 1:55(1983)
sprubps	rbcS	TAATTGGAAGATTTGAGA	190	38	f	PMB 7:433(1986)
strbcs1	rbcS	AACCACATATACATATCAAA	422	42	f	pers. comm. 1
strbcs1	rbcS	TCATACCCCAAGCATGGAT	2509	38	f	"
strbcs2a	rbcS	GATGGGAAGTTGAGTGAA	109	42	f	"
strbcs2a	rbcS	GCTTACCCCAAGCATGGAT	832	45	f	"
strbcs2b	rbcS	GGTGATCCCTTATGTGAA	29	41	f	"
strbcs2c	rbcS	AAGTCGGTTAGATATGAAG	2	37	f	"
strbcs2c	rbcS	GCTTACCCCAAGCATGGAT	702	45	f	"
strbcs-c13	rbcS	GCGTACCCCAAGCGCTGGAT	473	41	f	pers. comm. 2
strbcs-c52	rbcS	GCATACCCCAAGCAAGGAT	429	38	f	"
st-1s1	(unknown)	AGATATATTACTTGGTTGCA	558	37	f	MGG 205:14(1986)
st-1s1	(unknown)	aactccasgtgcacaaact	1353	33	r	"
tarub1	rbcS	GAGTACCGCTGAGCGGTACGT	229	38	f	NAR 11:8719(1983)
tarub2	rbcS	GAGTACCGCTGAGCGGTATGT	380	38	f	"
NON LIGHT INDUCED						
gmnod23	uricase	AGATGACCATCAAGTTCAAC	1159	39	f	NAR 13:239(1985)
maglinag	glutamine synthetase	GTCATGGTTGACAAGTGCA	3705	37	f	MGG 203:221(1986)
soy1bc	leghemoglobin	aattgttaactgtgtactctc	1094	33	r	NAR 10:689(1982)
consensus core sequence: (forward)						
(reverse)						

Figure 4. All sites found with the program EAGLE (17) which have a score greater than or equal to 37 in the case of the matrix LIGHT1.IMAT and 33 in the case of the matrix LIGHT1REV.IMAT. The position gives the first base of the LRE displayed in the corresponding sequence. The LREs found with LIGHT1.IMAT are written in capitals and indicated "f". The core motifs are inversed.

Declaration of abbreviations of reference sources: Bio = Biochimie; Book1 = Steinback, K.E., Bonitz, S., Arntzen, C.J. and Bogorad, L. (Eds), "Molecular Biology of the Photosynthetic Apparatus", Cold Spring Harbor Laboratory Press, NY, (1985), pp. 381-396; Btc = Biotechnology; CG = Current Genetics; EMBOJ = EMBO Journal; FEBSL = FEBS Letters; JBC = Journal of Biological Chemistry; JMAG = Journal of Molecular and Applied Genetics; MCB = Molecular and Cellular Biology; MGG = Molecular and General Genetics; NAR = Nucleic Acids Research; Nat = Nature; PMB = Plant Molecular Biology; PNAS = Proceedings of the National Academy of Sciences of the United States of America; pers. comm. 1 = Christian Fritz et al., MPI für Züchtungsforschung, personal communication; pers. comm. 2 = Frank Wolter et al., MPI für Züchtungsforschung, personal communication.

sequence name:	sequence encodes for:	reference:
<u>LIGHT INDUCED, POSITIVE REGULATED</u>		
amchs	chalcone synthase	MGG 202:429(1986)
atlhcp1	chloroA/B binding protein	NAR 14:4051(1986)
cornrbcs	ssu-RuBisCO	NAR 15:4360(1987)
crrbcs2	ssu-RuBisCO	JMB 191:421(1986)
cucumber	ssu-RuBisCO	
<u>LIGHT INDUCED, NEGATIVE REGULATED</u>		
asap5r	phytochrome	NAR 13:8543(1985)
<u>NON LIGHT INDUCED</u>		
atadh	alcohol dehydrogenase	PNAS 83:1408(1986)
chntleu	tRNA	CG 9:405(1985)
dcext	extensin	EMBOJ 4:2145(1985)
gmact1	actin	JMAG 2:111(1983)
gmbbi	protease inhibitor	JBC 259:9883(1984)
gmgly	glycinin	JBC 259:13436(1984)
gmglyl	glycinin	FEBSL 188:117(1985)
gmgl01	leghemoglobin	EMBOJ 2:449(1983)
gmgl05	leghemoglobin	PNAS 79:4055(1982)
gmhspgm	heat shock protein	PNAS 82:3726(1985)
gmhsp1	heat shock proten	EMBOJ 3:2491(1984)
gmhsp2	heat shock protein	"
gmhsp175	heat shock protein	MCB 5:3417(1985)
gm1bc3g1	leghemoglobin	NAR 13:239(1985)
gmlea	lectin	Cell 34:1023(1983)
gmnod35g	uricase	PNAS 82:5040(1985)
hvb1hord	hordein	EMBOJ 4:9(1985)
lep11	proteinase inhibitor	PNAS 83:7277(1986)
lewip1	protease inhibitor	JBC 260:6555(1985)
mibvmcp0	mitochondrial minicircle	NAR 12:4747(1984)
pslegag	legumin	NAR 12:4493(1984)
psvic7	vivilin	NAR 11:2367(1983)
pvdlec1	phytohemagglutinin	EMBOJ 4:883(1985)
pvllect	lectin	NAR 10:7819(1982)
soylba	leghemoglobin	NAR 10:689(1982)
soylbg11	leghemoglobin	PNAS 79:4055(1982)
tagligbc	gliadin	PMB 3:325(1984)
rcricin	ricin	NAR 13:8019(1985)
zmtac	transposable element AC	MGG 198:19(1984)

Figure 5. Sequences which obtained matrix scores below the threshold levels. For a list of abbreviations of the reference sources see figure 4.

generated. The matrix LIGHT1.IMAT, shown in figure 2, contains 23 sites found in forward (5' to 3') direction and the matrix LIGHT1REV.IMAT, shown in figure 3, contains 19 sites found in the reverse orientation. Both matrices contain in addition 20 bases upstream and downstream of each site.

		number of sequences found not found with LIGHT1.IMAT and LIGHT1REV.IMAT	
whole number of sequences	75	40	35
number of positive regu- lated light induced seq.	42 = 100%	37 = 88%	5 = 12%
number of negative regu- lated light induced seq.	1	0	1
number of non light induced sequences	32 = 100%	3 = 9%	29 = 91%

Figure 6. Results obtained searching 75 plant sequences with EAGLE (17) using the matrices LIGHT1.IMAT and LIGHT1REV.IMAT.

Sequence Analysis with Consensus Matrices

We searched 75 plant genes with the two matrices described above using the program EAGLE (17). By setting the threshold level of the matrix scores of LIGHT1.IMAT to 36 and of the matrix scores of LIGHT1REV.IMAT to 33 we detected 40 genes with matrix scores above the threshold levels. Of these 40, 24 represent the initial set from which the matrices were derived. The remaining 16 were identified from the database solely on the basis of matrix scores above the set thresholds. Of the 16 additional genes so identified, 13 are known to be under the control of the phytochrome system. The remaining three genes with sufficiently positive matrix scores are not known to be regulated by phytochrome. From a total of 42 phytochrome dependent light inducible genes in the database, 37 (88%) were identified at the described threshold levels and from the 32 investigated non light induced genes only 3 (9%) were predicted incorrectly to be light dependent. With the only negatively regulated light induced gene investigated (phytochrome) we obtained scores below the threshold levels with both matrices. Neither adding the two matrices together nor use of shorter partial matrices was able to give results better than the data presented (shorter matrices and respective data not shown). A compilation of the sites found is

given in figure 4. The sequences which contain only sites with scores below the threshold levels are listed in figure 5.

Since, with the methods employed, not only the majority of *rbcS* genes, but also genes encoding chlorophyll a/b binding proteins and other light inducible genes could be discerned from genes which are not light inducible, we conclude that a fundamental link exists between the presence of the consensus, which we suppose is a light responsive element (LRE), and the ability of the gene to be positively regulated at the transcriptional level in a phytochrome mediated response to light stimulus.

The results of the sequence analysis using consensus matrices are summarized in figure 6.

DISCUSSION

Comparison to Experimental Data

In a few cases experimental data for the sequences investigated here are available. By controlled deletion experiments attempts have been made to define LREs on the basis of functional analysis (21,22,23,24). It is now possible to compare these results with our predictions.

Figure 7 shows a segment of the sequence PSRC01, in which the predicted LREs have been marked. This sequence is called *rbcS-E9* in Morelli *et al.* (21). Morelli *et al.* concluded that in the PSRC01 gene the sequence from 1016 to 1049, which includes the TATA-box, contains the necessary information for light response. Though this apparently is in contrast to our results, inspection reveals that first a series of 5' deletions was used to find the 5' border of the LRE and in this experiment the constructs contained the entire 3' remainder of the gene. We predict two LREs residing in exon 3 and in the 3' untranslated region (see figure 6). To find the 3' border of the LRE Morelli *et al.* fused another gene to the PSRC01 promotor. Our analysis of this construct also reveals an LRE at position 859 to 897. In all constructs used in (21) we predict at least one LRE. Experiments have shown that LREs defined by other criteria (23,24) may act over a distance of at least a few hundred base pairs, regardless

Part of sequence PSRC01

```

-252
801  ttgtgttagt tttttcagtt agcttagtgg gcattttaca cgtggcattc
-202                                     score 38 reverse
851  ttatcctatt gctggctatt gatttggctta gcacacaaaa cttttcaatc
-152
901  ttgtgttggt aatatgactg caaagtttat cattttcaca atccaacaaa
-102
951  ctggttcttag gcagtggcca actaccacaa ttttaagacc ataatatgg
-52
1001 aaatagaaaa atcaatacat tatatatagc aagttttagt acaagctttg
-2
1051  tttattcaacc acaagaacta agaaagtcag aaaaaattgc TTCTATGATA
49
1101 TCCTCTTC CG CTGTGACAAC AGTCAGCCGT GCCTCTAGGG GGCAATCCGC
99
1151 CGCAGTGGCT CCATTGGCG GCCTCAAATC CATGACTGGA TTCCCACTGA
149
1201 AGAAGGTCAA CACTGACATT ACTTCCATTA CAAGCAATGG TGGAAGAGTA
199
1251 AAGTGCATGC AGgtgacaga aacatatata tatatatagt tgaatatcag
249
1301 taatgattca agttttgtta ccgttattgt tgaatattta gGTGTGGCCT
299
1351 CCAATTGGAA AGAAGAAGTT TGAGACTCTT TCCTATTTCG CACCATTGAC
349
1401 CAGAGATCAG TTGTTGAAAG AAGTTGAATA CCTTCTCAGA AAGGGATGGG
399
1451 TTCCTTGCTT GGAATTGAG TTGGAGgttt catattcatt ccttttttca
449
1501 atgattatat aaatactttt gtttgaaacc ctaatgagtt gattttgact
499
1551 gtttggttgc agAAAGGATT TGTGTACCGT GAGCACAACA AGTCACCAGG
549
1601 ATACTATGAT GGAAGATACT GGACAATGTG GAAGCTTCCT ATGTTTGGA
599
1651 CCACTGATGC TTCTCAAGTC TTGAAGGAGC TTGATGAAGT TGTTGCCGCT
649                                     score 39 forward
1701  TACCCCAAG CATTGCTTCG TATCATCGGT TGTGGAAGC TTCTCAAGT
699
1751 TCAATGCATC AGTTTCATTG CGCACACACC AGAATCCTAC TGAgttcgag
749
1801 tattatggca ttgggaaaac tgtttttctt gtaccatttg ttgtgcttgt
799
1851 aatttactgt gttttttatt cggttttcgc tatcgaactg tgaaatggaa
849                                     score 36 reverse
1901  atggatggag agaggttaat gattgatata gtccttttgt caatttcaa
899
1951  attaatatta ttgtttttt ctcttatttg ttgtgtgttg aatttgaaat

```

Figure 7. Part of sequence PSRC01 containing the predicted LREs (shaded, with inversed core motif). The TATA-box, the transcription start and the translation start are inversed. The numbers written in subscript denote the position according to the transcription start site.

of their orientation. Our results are therefore consistent with such data as obtained elsewhere (21).

Figure 8 presents segments of the sequence PSRBCS3A. Fluhr *et al.* (22) and Kuhlemeier *et al.* (23) performed experiments to

Part of sequence PSRBCS3A

```

-211                                     score 37 reverse
201  cacatggcac  tgtcctattg  gtggcttatg  ataaggctag  cacacaaaat
-161                                     Box II                               Box III
251  ttcaaatctt  gtgtggttaa  tatggctgca  aactttatca  ttttcaciat
-111
301  ctaacaagat  tggtagtagg  cagtagctaa  gtaccacaat  attaagacca
-61
351  taatattgga  aatagataaa  taaaaacatt  atatatagca  agtttttagca
-11
401  gaagctttgc  attcatatac  gaagtggaaa  aaATGGCTTC  TATGATATCC
40
451  TCTTCAGCTG  TGACTIONAGT  CAGCCGTGCT  TCTACGGTGC  AATCGGCCGC
90
501  GGTGGCTCCA  TCGGCGGCC  TCAAATCCAT  GACTGGATTG  CCAGTTAAGA
140
551  AGGTCAACAC  TGACATTACT  TCCATTACAA  GCAATGGTGG  AAGAGTAAAG
190
601  TGCATGCAGg  tcagaaaaat  atatattctc  ttgtataaa  ctttttggtt
240
651  ctgtttgatt  ggtttatttt  ggtgtatcta  ttggcataag  cacttattag
290                                     score 33 reverse
701  accgcttgag  agaggtttag  gaaaccaatt  atgataagtc  cataaacggt
340
751  ttcaacttat  ttttgtatgt  tcttcaggtt  aggttatgaa  aataacttat

////////////////////////////////////
990
1401  TCTCAAGTCT  TGAAGGAGCT  TGATGAAGTT  GTTGCOCGCT  ACCCCAAGC
1040                                     score 42 forward
1451  TTTCGTCGGT  ATCATCGGTT  TCGACAACGT  TCGTCAAGTT  CAATGCATCA
1090
1501  GTTTCATTGC  CCACACACCA  GAATCCTACT  AAgtttgagt  attatggcat
1140
1551  tggaaaagct  gttttcttct  atcatttggt  ctgcttgtaa  tttacttgtt
1190
1601  tctttcagtt  tttgttttcg  gacatcaaaa  tgcaaatgga  tggataagag
1240                                     score 36 reverse
1651  ttaataaatg  atatggtcct  tttgttcatt  ctcaaatatt  tattatctgt

```

Figure 8. Part of sequence PSRBCS3A containing the predicted LREs (shaded, with inversed core motif). The boxes II and III of (23), the TATA-box, the transcription start and the translation start are inversed. The numbers written in subscript denote the position according to the transcription start site.

locate the LREs in this gene. They concluded that the 5' end of a positive acting LRE is located between -166 and -149 (relative to the transcription start) and extends beyond -100 (23). They also suppose that there may be multiple LREs in one gene and that a second positive LRE most likely is present in the far upstream region between -410 and -170. We obtained 4 scores above the threshold level in this sequence (figure 4), one of them residing between -205 and -156. Furthermore Kuhlemeier *et al.* located an element which influences transcription negatively in light dependent manner between -169 and -112 (23).

Part of sequence PSRBCS36

```

-224                                     score 38 reverse
751  aatgggcatc ttacacgtgg cattatccta ttgggtggcaa atgataaggc
-174
801  agggaacacac aacttttcaa tcttgtgtgg ttaatatggc tgcaaagttt
-124
851  atcattttcac aatctaacaa gattgtgtact aggcagtagc taattaccac
-74
901  aatattaaga ccataatatt ggaaatagat aaataaaaac attatatata
-24
951  gcaagtttta gcagaagcctt tggaattcat acagaagtga gaaaaATGGC
27
1001 TTCTATGATA TCCTCTTCCG CTGTGACAAC AGTCAGCCGT GCCTCTAGGG
77
1051 GGCAATCCGC CGCAGTGGCT CCATTGGGCG GCCTCAAATC CATGACTGGA
127
1101 TTCCCAAGTGA AGAAGGTCAA CACTGACATT ACTTCCATTA CAAGCAATGG
177
1151 TGGAAGAGTA AAGTGCATGC AGgtgacaga aacatatata tatatatata
227
1201 tagttgaata tcagtaatga ttcaagtttg ttaaccgttt atgttgaata
277
1251 tttagGTGTG GCCTCCAATT GGAAAGAAGA AGTTTGAGAC TCTTTCCTAT
327
1301 TTGCCACCAT TGACGAGAGA TCAATTGTTG AAAGAAGTTG AATACCTTCT
377
1351 GAGGAAGGGA TGGGTTCCAT GCTTGAATT TGAGTTGGAG gtttcatatt
427
1401 cattcctttt ttcaatgatt atataaatac ttttgttga aaccgtaatg
477
1451 agttgatttt gactgttttg ttgcagAAAG GATTGTGTGA CCGTGAGCAC
527
1501 AACAAAGTCAC CAGGATACTA TGATGGAAGA TACTGGACAA TGTGGAAGCT
577
1551 TCCTATGTTT GGTACCACTG ATGCTTCTCA AGTCTTGAAG GAGCTTGATG
627                                     score 39 forward
1601 AAGTTGTTGC CGCTTACCGT CAGGCTTCCG TCGTATCAT CGGTTTCAC
677
1651 AACGTTCCGT AAGTTCAATG CATCAGTTTC ATTGCACACA CACCAGAATC

```

Figure 9. Part of sequence PSRBCS36 containing the predicted LREs (shaded, with inversed core motif). The TATA-box, the transcription start and the translation start are inversed. The numbers written in subscript denote the position according to the transcription start site.

Figure 9 shows a segment of the sequence PSRBCS36 (named *rbcs* ss3.6 in ref. 24) which was purported to contain an LRE (24). In analysis of 5' deletion constructs, the 5' border of the putative LRE in the system under study was assigned to position -92 (corresponding to position 880 in figure 9). Yet a construct containing this region yielded a conversion difference between light and dark of only 2% (plasmid pMH-92; conversion dark: 1.8%; conversion light: 3.8%). This value lies precariously close to the limit of solution of the method employed, since another construct which is not light inducible has a conversion differ-

ence of 1.6% (plasmid pNCAT-4; conversion dark: 5.6%; conversion light: 4.0%). The next 5' deletion construct reaches up to -357 and gives a conversion difference of 3.6%. We suggest that the first LRE of this gene lies within position 771 to 819 (corresponds to -201 to -152 of Timko *et al.*). A second LRE in this gene is detected in exon 3.

It is very difficult to directly compare the experimental data obtained by independent groups in this field of genetics due to the lack of stringently defined and controlled gene transfer and gene product quantification methods. The extrapolation of our analytical results onto experimental data must necessarily be viewed in light of this circumstance. So far no one has definitely identified an LRE of the type we propose. Our method enables us to make predictions. Such predictions, derived from analytical comparisons, are thus far in good agreement with the available experimental data relevant to LREs.

Possible Genetic Elements Responsible for Phytochrome Dependent Light Induction and their Arrangement within a Gene

A complex pattern for phytochrome dependent light induction can be derived from the literature. Serfling *et al.* (25) pointed out that enhancers generally have a modular structure with highly redundant information in which conserved motifs of several kinds are separated by less conserved DNA sequences. It has been demonstrated that with a set of increasingly extensive 5' deletions the light induction of *rbcS* genes can be decreased stepwise (21,24). Kuhlemeier *et al.* (23) showed that even if they deleted the 5' region of the *PSRBCS3A* gene up to -100, this construct still showed a low level of light dependent transcription. Therefore, they suppose that there may be an additional LRE downstream from -100 in *PSRBCS3A*.

The following model tries to reconcile all these experimentally found pieces of the puzzle: Due to the fact that we can find the LRE described in this paper in almost all phytochrome dependent regulated genes we suppose that this element is a docking motif of a hypothetical light dependent amplification mediating protein (LAMP) which transmits the light stimulus to the genes. This LAMP binding site is part of a modular enhancer pattern but itself can not enhance significantly the

transcription level. The LAMP binding sites can be spread over the whole gene and mostly they are redundant within a single regulated gene. The binding site is orientation independent and has the core consensus sequence "CCTTATCAT". There may be further enhancer blocks within the introns and exons. This regulatory pattern can also include silencer blocks (23,26). Additional overlapping regulatory elements, in example motifs controlling tissue specific expression, are supposed by experimental data from *rbcS* and chlorophyll *a/b* binding protein genes (22,26,27,28,29,30). It is important to notice, that, derived from the existing experimental and computational data, it is quite sure, that the enhancers are independent from the positive and negative controlling LREs. This fact has a great influence on the design of experiments to localize such LREs.

It is also possible, that the found sequence motif may serve as a constitutive enhancer within these genes.

If one has a gene with known numbers of positive and negative acting LREs (for example *PSRBCS3A*), one can modify selectively the known LREs (in our example Box II and Box III) and/or the positive acting LREs proposed in this paper. These modifications can be done by using synthetic oligomers with altered sequences exchanging the synthesized oligomers with the LREs. Switching off selectively the different boxes while leaving others intact leads to a response pattern from which the roles of the different elements may be elucidated.

Subtypes of LAMP Binding Sites

As shown in figure 4 the gene of chalcone synthase investigated here is predicted to be not light induced. In (9) it is shown that this particular gene is light induced. Because chalcone synthases are used in a different metabolic pathway than *rbcS* gene products (31,32) it is likely that the light dependent regulation is somewhat different. One possibility to mediate different light stimuli is to alter slightly the LAMP and its corresponding binding sites. This would give at least three different subtypes of LAMP binding sites derived from a common ancestor, one for each light receptor. This can be proven when LAMP binding sites of all three light-stimulus mediating signal systems become known.

Future Improvements in Recognizing this Type of LREs

For the first time we present a sequence motif common to almost all phytochrome dependent light inducible plant genes. This motif can therefore be used to discriminate these genes from non light induced plant genes with a high probability (greater than 85%). We call this sequence element LAMP binding site. To improve the accuracy of the method it is necessary to identify some of the proposed LREs experimentally. With such informations we will be able to create better and shorter matrices and merge our two different matrices.

ACKNOWLEDGMENTS

The authors are indebted to Christian Fritz, Frank Wolter and Dr. Peter Schreier for using their nucleic acid sequences of different rbcS genes prior to publication. We thank William Martin, Dr. Peter Schreier and Frank Wolter for critical reading of the manuscript. U.G. was supported by a Ph.D. fellowship from the Max Planck Society.

*To whom correspondence should be addressed

REFERENCES

1. Tobin, E.M. and Silverthorne J. (1985) *Ann. Rev. Plant Physiol.* 36, 569-593.
2. Pratt, L.H. (1979) *Photochem. Photobiol. Rev.* 4, 59-124.
3. Tobin, E.M. (1981) *Plant Mol. Biol.* 1, 35-51.
4. Thompson, W.F., Everett, M., Polans, N.O., Jorgensen, R.A. and Palmer, J.D. (1983) *Planta* 158, 487-500.
5. Silverthorne, J. and Tobin, E.M. (1984) *Proc. Natl. Acad. Sci. USA* 81, 1112-1116.
6. Apel, K. (1979) *Eur. J. Biochem.* 97, 183-188.
7. Mössinger, E., Batschauer, A., Schäfer, E. and Apel, K. (1985) *Eur. J. Biochem.* 147, 137-142.
8. Apel, K., Gollmer, I. and Batschauer, A. (1984) *J. Cell Biochem.* 23, 181-189.
9. Kaulen, H., Schell, J. and Kreuzaler, F. (1986) *EMBO J.* 5, 1-8.
10. Thien, W. and Schopfer, P. (1982) *Plant Physiol.* 69, 1156-1160.
11. Mössinger, E. and Schopfer, P. (1983) *Planta* 158, 501-511.
12. Gallagher, T.F. and Ellis, R.J. (1982) *EMBO J.* 1, 1493-1498.
13. Otto, V., Mössinger, E., Sauter, M. and Schäfer, E. (1983) *Photochem. Photobiol.* 38, 693-700.
14. Hershey, H.P., Barker, R.F., Idler, K.B., Lissemore, J.L. and Quail, P.H. (1985) *Nucleic Acids Res.* 13, 8543-8559.

15. Devereux, J., Haeberli, P. and Smithies, O. (1984) *Nucleic Acids Res.* 12, 387-395.
16. Grob, U. and Stüber, K. (1987) *Comp. Appl. Biosci.* 3, 243-244.
17. Grob, U. and Stüber, K. (1988) joint issue of *Nucleic Acids Res.* and *Comp. Appl. Biosci.* in press.
18. von Hippel, P.H. (1979) in: Goldberger (Ed), *Regulation and Development* 1 pp.279-347.
19. Bucher, P. and Bryan, B. (1984) *Nucleic Acids Res.* 12, 287-305.
20. Bucher, P. and Trifonov, E.N. (1986) *Nucleic Acids Res.* 14, 10009-10026.
21. Morelli, G., Nagy, F., Fraley, R.T., Roger, S.G. and Chua, N-H. (1985) *Nature* 315, 200-204.
22. Fluhr, R., Kuhlmeier, C., Nagy, F. and Chua, N-H. (1986) *Science* 232, 1106-1112.
23. Kuhlmeier, C., Fluhr, R., Green, P.J. and Chua, N-H. (1987) *Genes and Development* 1, 247-255.
24. Timko, M.P., Kausch, A.P., Castresana, C., Fassler, J., Herrera-Estrella, L., Van den Broeck, G., Van Montagu, M., Schell, J. and Cashmore, A.R. (1985) *Nature* 318, 579-582.
25. Serfling, E., Jasin, M. and Schaffner, W. (1985) *Trends in Genetics* 224-230.
26. Simpson, J., Schell, J., Van Montagu, M. and Herrera-Estrella, L. (1986) *Nature* 323, 551-554.
27. Coruzzi, G., Broglie, R. Edwards, C. and Chua, N-H. (1984) *EMBO J.* 3, 1671-1679.
28. Dean, C., van den Elzen, P., Tamaki, S., Dunsmuir, P. and Bedbrook, J. (1985) *EMBO J.* 4, 3055-3061.
29. Simpson, J., Timko, M.P., Cashmore, A.R., Schell, J., Van Montagu, M. and Herrera-Estrella, L. (1985) *EMBO J.* 4, 2723-2729.
30. Simpson, J., Van Montagu, M. and Herrera-Estrella, L. (1986) *Science* 233, 34-38.
31. Kreuzaler, F. and Hahlbrock, K. (1975) *J. Biochem.* 56, 205-213.
32. Miziorko, H.M. and Lorimer, G.H. (1983) *Ann. Rev. Biochem.* 52, 507-