Using Cauliflower to Find Conserved Non-Coding Regions in Arabidopsis[1]

Juliette Colinas, Kenneth Birnbaum, and Philip N. Benfey*

Department of Biology, 1009 Main Building, New York University, 100 Washington Square East, New York, New York 10003

A bioinformatics approach is used to analyze the degree of conservation between upstream non-coding regions of cauliflower (*Brassica oleracea*) and Arabidopsis. The level of homology suggests that comparison of these two species could reveal functional cis-regulatory elements.

There is growing interest in comparing genome sequences to identify regulatory regions (Stojanovic et al., 1999). This arises in part from the failure of de novo computational methods to consistently recognize functional promoter elements from single genomes (Loots et al., 2000; Pennacchio and Rubin, 2001). Because genomic regions that have a biological function are often conserved through evolution, non-coding regions conserved between species are more likely to contain regulatory sequences (Stojanovic et al., 1999). Numerous computer programs have been written to extract conserved regions or motifs from orthologous sequences (reviewed elsewhere; Fickett and Wasserman, 2000; Stormo, 2000; Ohler and Niemann, 2001). In addition, several studies have shown that the conserved non-coding sequences (CNS) found using such comparisons often have biological meaning (Hardison, 2000; Kent et al., 2000; Loots et al., 2000) and are enriched in transcription factor binding sites (Levy et al., 2001).

The genomes to be used in these comparisons must be carefully selected if useful results are to be obtained; comparison of too closely related genomes identifies nonfunctional conservation, whereas too distantly related genomes lack sufficient conservation for a meaningful comparison. Evidence from studies in animals and bacteria suggest that more closely related species are more likely to be useful for identification of regulatory regions because they appear to change more rapidly than coding regions (Huynen and Bork, 1998; Cargill et al., 1999).

Among plants, extensive genomic sequence is at present only available for Arabidopsis. As a consequence, the choice of additional plant species to sequence is important to provide maximal information from sequence comparisons. This choice could be made if sequence data were available from a number of related plant species, but presently limited sequence data is only available for cauliflower. The genus *Brassica* includes many species and cultivars (O'Neill and Bancroft, 2000) for which there are economical incentives for genome sequencing. This genus is closely related phylogenetically to Arabidopsis, their divergence time being estimated at 14.5 to 20.4 million years based on mitochondrial DNA data (Quiros et al., 2001). However, it is still unclear how much conservation can be found in the non-coding genomic regions of these two genera. In an analysis of the promoter of *APETALA3* orthologs in Arabidopsis and cauliflower, Hill et al. (1998) found 62% identity in the 440 bases upstream of the transcription start site. However, another study comparing a genomic region between cauliflower and Arabidopsis found less identity in several promoter comparisons, except for one region of 59 bp with 78% identity in one promoter and a 340-bp region with 54% identity in another promoter (Quiros et al., 2001). Thus, it is important to expand upon such analyses to establish whether a comparison of Arabidopsis with cauliflower is likely to provide useful regulatory site information. The study described here is a first step toward answering that question. Using more extensive data now available for cauliflower from a shotgun-sequencing project along with the completed Arabidopsis sequence, we conducted a preliminary comparison of cauliflower and Arabidopsis putative regulatory regions.

Cauliflower shotgun sequences (8,864 total) of about 400 to 700 bp in length (covering about one-hundredth of the estimated 600-Mb genome; O'Neill and Bancroft, 2000) were obtained from Washington University and Cold Spring Harbor Laboratory (ftp://cshl.org/pub/sequences/brassica_shotgun/, submitted on 2001/05/04) and were subjected to a BLAST analysis (http://www.ncbi.nlm.nih.gov/BLAST/; Altschul et al., 1997) against the entire National Center for Biotechnology Information nucleotide database, including expressed sequence tags. To identify the best candidate sequences for comparative analysis, a program was written to select the cauliflower sequences that were homologous to the 5′ end of an Arabidopsis gene and also contained part of the 5′ non-coding region of that gene. This was done by screening the BLAST output to select for

[1] This work was supported by the National Institutes of Health (grant no. GMR01–43788 to P.N.B.).

* Corresponding author; e-mail philip.benfey@nyu.edu; fax 212–995–4204.

www.plantphysiol.org/cgi/doi/10.1104/pp.002501.

cauliflower sequences that hit at least one non-plastid and non-ribosomal complete Arabidopsis cDNA with an alignment of at least 50 bp and an overhang at the 5′ end of the cDNA of at least 100 bp. To ensure that true orthologs were compared, the alignments of the 60 cauliflower and Arabidopsis sequences retrieved from this first selection were then manually inspected. Only the cauliflower shotgun sequences that aligned with a BLAST score above 80 to a single Arabidopsis genomic fragment which also aligned almost perfectly with the originally identified Arabidopsis cDNA were kept for further analysis. Twenty-six sequences were rejected based on these criteria. In addition, 13 sequences were discarded due to inconsistent annotation in Arabidopsis (that is, the cDNA annotation contradicted the genomic annotation). Finally, eight sequences aligned to Arabidopsis ribosomal or plastid cDNA that were not annotated as such. Thirteen of the initial 60 cauliflower sequences were, thus, kept and analyzed further. Because we selected only promoters with the best evidence for orthology between cauliflower and Arabidopsis, it is probable that more cauliflower sequences could have been analyzed, but we decided to use conservative criteria.

The 13 cauliflower shotgun sequences retained from the manual selection were aligned with Arabidopsis using VISTA (http://www-gsd.lbl.gov/vista/; Mayor et al., 2000; Dubchak et al., 2000), which can find windows of high identity in an alignment that shows generally poor conservation. Because some of the 5′ non-coding regions compared were around 100 bp and we wanted to identify small regions of conservation, a window size of 25 bp was chosen. Eight

Table I. *Summary of the VISTA alignment results between 13 cauliflower shotgun sequences and their Arabidopsis homologs*

Gene abbreviations: RSH3, RelA/SpoT homolog 3; PLC, phospholipase C; SIMIP, salt-stress-inducible major intrinsic protein; Syp42, syntaxin of plants 42; and DRT112, recombination and DNA damage resistance 112. The "unknown proteins" are unidentified sequenced cDNA clones.

	Cauliflower Shotgun Sequence Identifier[a]	Putative Arabidopsis Ortholog[b]	Accession No.[c]	Approximate 5′ Non-Coding Overlap (bp)[d]	Length and Position of Conserved Non-Coding Sequences			
					5′ Non-coding[e]		5′-Untranslated region[f] (when identified)	
					Size of CNS	Distance from translation start site	Size of CNS	Distance from translation start site
1	ef15d07.x1	Histone H1-3	U73781.1	260	0		0	
2	ef18a02.x1	RSH3	AF225704.1	301	65	207	47	171
					30	143		
3	ef20b10.x1	DRT111	M98455.1	200			43	97
					35	56		
4	ef21a02.x1	Nicotianamine synthase	AB021934.1	540	66	150	N/A	
					28	119		
5	ef28h04.x1	Immunophilin	AF370159.1	350			0	
					31	288		
6	ef29c10.x1	Zinc finger protein 2	AF138744.1	130	59	50	N/A	
					56	378		
7	ep72h01.b1	PLC	AF360206.1	450			0	
					33	329		
					48	98		
8	jnr37a04.b1	SIMIP	AF003728.1	450			N/A	
					80	15		
					25	142		
9	jnr52e04.b1	Syp42	AF154574.1	180			N/A	
					81	52		
10	jnr21f10.g1	DRT112	AF361853.1	240	0		0	
					48	280		
					67	215		
11	ef29d04.x1	Unknown protein 1	AF375417.1	530	29	141	46	0
					34	107		
					44	56		
							25	52
12	ef29h05.x1	Unknown protein 2	AF332423.1	410	118	112		
							47	71
13	jnr65c10.b1	Unknown protein 3	AF387005.1	200	0		0	

[a] Sequence identifier given by CSHL or Washington University. [b,c] Gene name and cDNA accession number of the identified Arabidopsis ortholog. [d] Approximate length available for comparison 5′ to the translation start site, in bp. [e] Size of conserved non-coding regions (defined here as a region of at least 25 bp with at least 75% identity) and the distance in bp between the translation start site and the 3′ of the CNS (i.e. the end closest to the start site), as identified by the VISTA plots and listings of the conserved regions. [f] For the genes for which annotation of the 5′-untranslated region is available, the CNSs that fall in this region are indicated in the last column.

negative controls using random pairs of cauliflower and Arabidopsis non-coding sequences revealed that windows of 25 bp with at least 75% identity were unlikely to occur by chance alone (none was found in the eight random comparisons). A CNS was, thus, defined here as having 75% or more identity over a window of at least 25 bp.

The alignments obtained for the 13 sequences are summarized in Table I and two representative alignments are illustrated in Figure 1. The results show that 10 of the 13 genes contain at least one conserved region between cauliflower and Arabidopsis in their 5′non-coding sequence. The size of these regions varies between 25 and 118 bp and averages 48 bp, and 37% of the non-coding bases belong to a conserved region. Most genes (8/10) contain one to two CNSs, which are always separated either from the site of translation initiation (as in the RSH3 gene of Fig. 1) or transcription initiation (as in the unknown gene of Fig. 1) by a region of low conservation of at least 30 bp. As seen in Table I (column "distance from translation start site"), CNSs can be found at distances from the translation start ranging from 46 to 434 bp, whereas the sizes of the non-coding regions available for comparison range from 130 to 540 bp. Thus, CNSs can be found throughout the non-coding sequences. Although it is possible that some of these CNSs represent cryptic exons, this is unlikely to be the case for all the genes compared. We also note that for one of the three genes for which no CNS is found (unknown protein 3), the coding sequence conservation is poor (150 bp of 500), indicating that the two sequences might not be true orthologs because all the other genes show almost complete conservation in the coding region available for comparison. Finally, no CNSs were found in the introns that were available for comparison (three sequences).

Because most sequence comparison analyses have been carried out between much more distantly related animal or bacterial species, e.g. mouse and human, which are separated by about 80 million years (Hardison et al., 1997), one question was whether there would be too much conservation between Arabidopsis and cauliflower for most of these

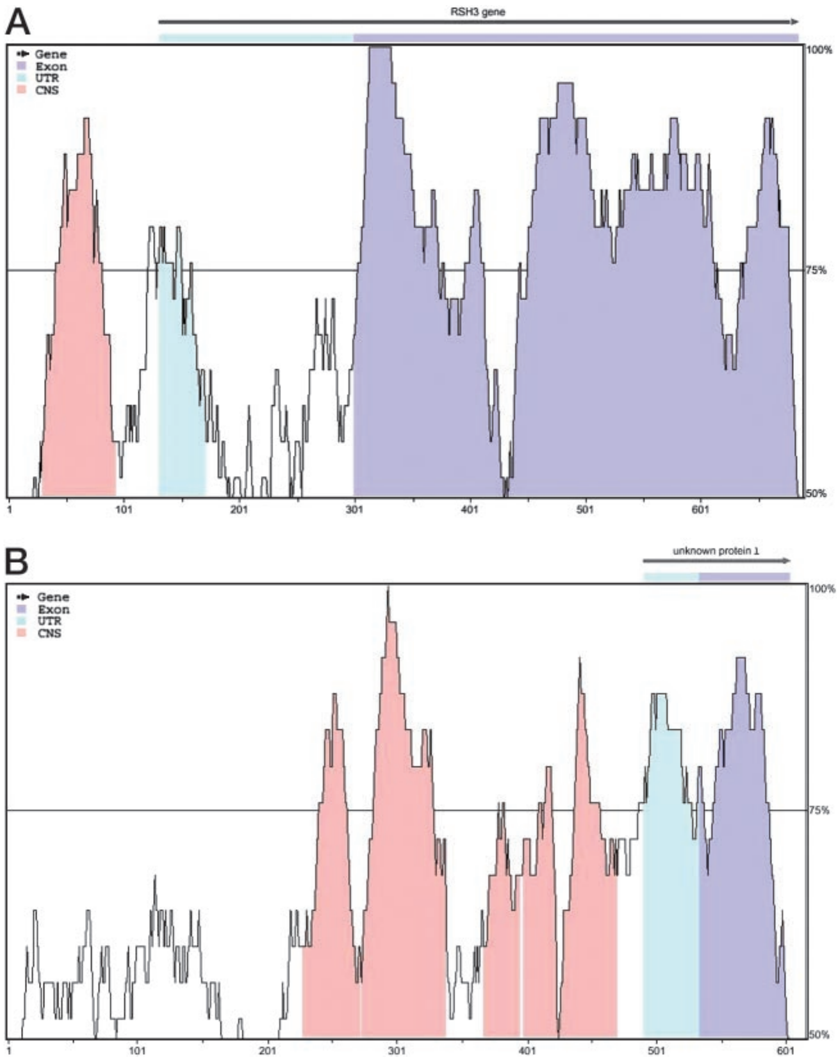

Figure 1. Examples of VISTA alignments of cauliflower shotgun sequences with their Arabidopsis homologs. The alignments for the RSH3 gene (top) and unknown protein 1 (bottom) are shown. The horizontal and vertical axes represent the position in the sequences (in basepairs), and the percent identity of the two sequences in a 25-bp window around that position, respectively. Regions in which the identity is greater than or equal to 75% are colored in pink (for non-coding regions), turquoise (5′-untranslated region [UTR]), or blue (coding region). The level of conservation observed in the coding region and the short, relatively well-defined region of conservation in the non-coding region is representative of most of the others genes examined.

CNSs to be functionally meaningful. However, the degree of conservation of non-coding sequences does not seem to be greater than between mice and human. Levy et al. (2001) found that 20% of the bases in the upstream 500 bp of 502 disease genes from human and mouse are aligned by BLAST (parameters: match = +1 and mismatch = −1). Performing a similar analysis with our sequences, we also find an average of 20% conservation (data not shown). This number might be an overestimate because we are comparing shorter sequences and most of the conservation might be expected to lie proximal to the 5′ end of the genes, but it shows that the level of conservation between Arabidopsis and cauliflower does not seem to be dramatically higher than between mouse and human. Nevertheless, the functional significance of these CNSs remains to be experimentally tested.

Overall, even though the comparison set is small, this study indicates that there is likely to be significant conservation of promoter regions between Arabidopsis and cauliflower. This suggests that sequence comparisons across these two species may prove useful for the identification of regulatory regions. Coupled with experimental studies, conducting similar pilot studies with other plant species would allow the identification of the most informative plant species for sequence comparison with Arabidopsis.

ACKNOWLEDGMENTS

We thank Richard McCombie from Cold Spring Harbor Laboratory for providing the cauliflower shotgun sequences, Dennis Shasha for discussion, and Mike Chou and Borislav Iordanov for help with the programming.

Received January 9, 2002; accepted March 17, 2002.

LITERATURE CITED

Altschul SF, Madden TL, Shäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acid Res **25:** 3389–3402

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N et al. (1999) Nat Genet **22:** 231–238

Dubchak I, Brudno M, Loots GG, Mayor C, Pachter L, Rubin EM, Frazer KA (2000) Genome Res **10:** 1304–1306

Fickett JW, Wasserman WW (2000) Curr Opin Biotechnol **11:** 19–24

Hardison RC (2000) Trends Genet **16:** 369–372

Hardison RC, Oeltjen J, Miller W (1997) Genome Res **7:** 959–966

Hill T, Day CD, Zonslo SC, Thackeray AG, Irish VF (1998) Development **125:** 1711–1721

Huynen MA, Bork P (1998) Proc Natl Acad Sci USA **95:** 5849–5856

Kent WJ, Zahler AM (2000) Genome Res **10:** 1115–1125

Levy S, Hannenhalli S, Workman C (2001) Bioinformatics **17:** 871–877

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000) Science **288:** 136–140

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000) Bioinformatics **16:** 1046–1047

Ohler U, Niemann H (2001) Trends Gen **17:** 56–60

O'Neill CM, Bancroft I (2000) Plant J **23:** 233–243

Pennacchio LA, Rubin EM (2001) Nat Rev Genet **2** 100–109

Quiros CF, Grellet F, Sadowski J, Suzuki T, Li G, Wroblewski T (2001) Genetics **157:** 1321–1330

Stojanovic N, Florea L, Riemer C, Gumucio D, Slightom J, Goodman M, Miller W, Hardison R (1999) Nucleic Acids Res **27:** 3899–3910

Stormo GD (2000) Bioinformatics **16:** 16–23