

1P-C-223

Analysis of Relationship between Upstream Element of Genes and Expression Data Profile from Microarray Analysis: An attempt to Find Cis-elements by Combining Method of Motif Searching and Data Mining

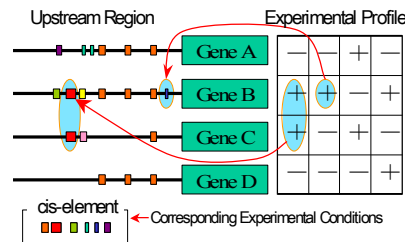
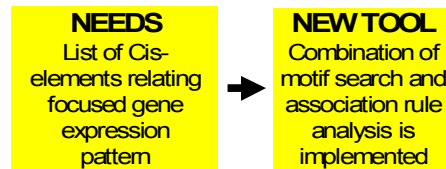
○Koji Doi¹, Toshifumi Nagata¹, Kouji Satoh¹, Kohji Suzuki², Shigemi Iizumi¹,
Setsuko Kimura¹, Aeni Hosaka¹, Shoshi Kikuchi¹

¹National Institute of Agrobiological Sciences, ²Hitachi Software Engineering

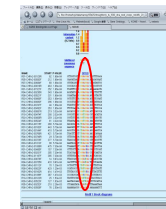
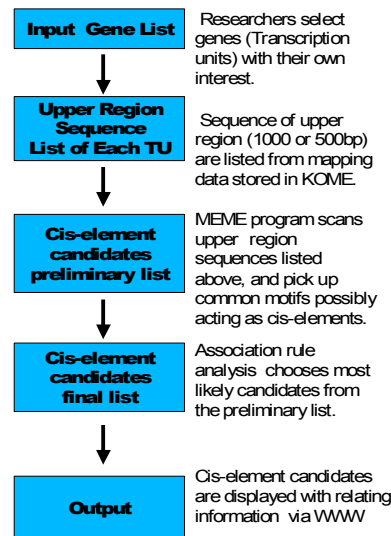
1. Introduction

- Accumulated information about over 30,000 full-length cDNA and microarray gene expression data of *Oryza sativa* enabled us to find motifs commonly existing beside genes simultaneously expressing. This process is important to find out regulatory gene networks, because these motifs are expected to play key roles in these networks, and it also suggests the existence of key trans elements.
- We are developing a data mining tool to find cis-element candidates from gene lists defined by researchers. Here we report the outline of the tool and the result of preliminary biological test of its usefulness.

2. Objective: Clarification of the Regulatory Network of Gene Expression



3. Flowchart of the Analysis



Common motifs among TUs listed in the first step were searched using MEME program.

<http://meme.sdsc.edu/meme/>

5. Association Rule Analysis

Reliability of relationships can be quantitatively evaluated by indexes listed below.

Rule = "If X occurs, then Y occurs."
If Y is not frequent in the whole but frequently occur with X, X and Y must be related.

Support	The value dividing the number of transactions containing both X and Y, by the total number of transactions.
Expected Confidence	The frequency of the number of transaction containing Y, in all transactions.
Confidence	The frequency of Y in transactions containing X. Ratio of support and Expected Confidence.
Lift	Ratio of Confidence to Expected confidence.

Example:

		Y		Total	
		Yes	No		
X	Yes	2	5	7	Support $\frac{2}{10} = 0.2$
	No	0	3	3	Confidence $\frac{2}{2} = 1.0$
Total		2	8	10	Lift $\frac{2/2}{7/10} = 1.43$

High lift value suggests strong relationship between X and Y.

6. Verification of the Result using Information of AtcisDB in AGRIS

NAME	SEQUENCE	OTHER INFO
ER7	TGTCTCCCAAAGGGAGACA	palindrome
DR5	TGTCTCCCTTTTGTCTC	direct repeat
DR5	GAGACA...AAAGGGAGACA	direct repeat inverse

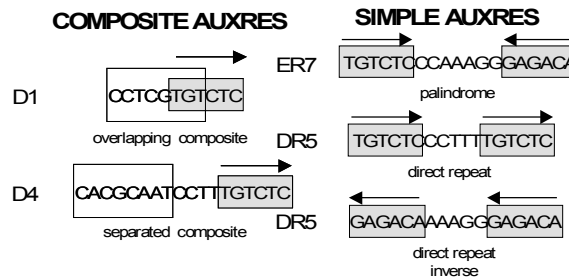
Cis-element candidates are verified with binding site motifs listed in AtcisDB.

• <http://arabidopsis.med.ohio-state.edu/AtcisDB/>

7. Biological Interest of Cis-Element Interaction

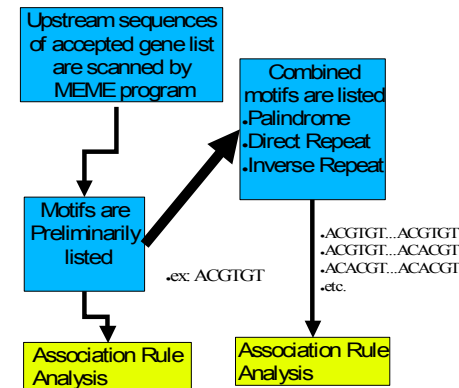
•Some cis-elements need to interact with other element to function. This is important phenomena to understand the gene expression regulatory mechanism of cis-elements.

•One good example is a cis-element called "AuxRE" concerned in auxin-responsive relation.



8. Simple Motif Search to Combined Motif Search:

Attempt to Find Cis-Elements Functioning in Combination



9. Case Study: Auxin Responsive Genes

•Biologically important genes relating to plant growth, development, etc.

•A cis-element 'AuxRE' is reported to be participant to regulate expression of auxin responsive genes (Guilfoyle et al. 1998).

- TGTCTC motif in AuxRE must interact with its palindromic element or multimerize to regulate gene expression.
- Several cis-elements with similar sequence of AuxRE are also known.

10. Material and Method

Gene List 1: ARF/IAA genes of *Arabidopsis thaliana* stored in RiceTFDB (2.0) (<http://ricetfdb.bio.uni-potsdam.de/v2.0/>) :

ARF/IAA genes themselves are auxin-inducible. Blastn search with these sequences to GenBank resulted 28 corresponding rice transcription units (TUs).

Gene List 2: KOME database includes 134 records of TUs relating to GO:009733="response to auxin".

MEME: Motif length: 6,7 and 8 bases, with 'Oops' and 'Zoops' options. MEME program was repeatedly called with 6 combinations of these parameters.

Generation of combined motifs: Combined motifs were made from motifs listed by MEME along three combination patterns (palindrome, direct repeat, and invert direct repeat) with flexible length (0-50bp) of gap.



11. Result (1)

Gene list 1:

MEME listed 5246 motifs from their upstream sequences (1000bps), of which 4579 TUs showed high lift value (>1.0).

There were 16 TUs showing partial match to the ATCIS records of "PRHA binding sites", which is described as "Developmental and auxin-induced expression of the *Arabidopsis* prha homeobox gene" (Table 1).

There was one motif matched to "ARF1 binding site", of which sequence is same as that of "AuxRE".

Possible 15738(=5246x3) combined sequence pattern were searched on upstream sequence of listed genes, and 1103 of them were found. Table 2 shows 50 motifs of them showing best lift values.

Table 1: Cis-element candidate motifs corresponding to AUX/IAA genes and suggested to be auxin-induction related according to ATCIS.

Motif	Hit TU in target group *1	Hit TU in the whole *2	Support	Confidence	Lift	ATCIS Description *3
ACACAC	7	2873	0	0.25	1.802	PRHA BS in PAL1
ACATTA	10	3232	0	0.357	2.289	PRHA BS in PAL1
ACATTAT	4	1190	0	0.143	2.487	PRHA BS in PAL1
ACTCAA	4	2842	0	0.143	1.041	PRHA BS in PAL1
ATACAC	7	2446	0	0.25	2.117	PRHA BS in PAL1
ATACACAC	3	347	0	0.107	6.396	PRHA BS in PAL1
ATACATT	3	1260	0	0.107	1.761	PRHA BS in PAL1
CAATTA	6	3775	0	0.214	1.176	PRHA BS in PAL1
TACACA	6	3148	0	0.214	1.41	PRHA BS in PAL1
TACATT	7	3842	0	0.25	1.348	PRHA BS in PAL1
TACATTA	3	993	0	0.107	2.235	PRHA BS in PAL1
TATACA	10	3802	0	0.357	1.946	PRHA BS in PAL1
TATACACA	2	351	0	0.071	4.215	PRHA BS in PAL1
TGTCTC	4	2088	0	0.143	1.417	ARF1 binding site motif
TTATACAC	1	202	0	0.036	3.662	PRHA BS in PAL1

*1 The number of TU possessing the designated motif within 28 TUs of the target gene list.
 *2 The number of TU possessing the designated motif within 20648 TUs stored in KOME database.
 *3 ARF1=Dimerization and DNA binding of auxin response factors
 PRHA=Developmental and auxin-induced expression of the Arabidopsis prha homeobox gene

Table 2: Top 50 motifs in lift values for ARF/IAA genes.

Motif	Hit TU in target group	Hit TU in the whole	Support	Confidence	Lift
TCTTTCAC[ACGT]{0.50}GTGAAAGA	1	1	0.000	0.036	739.786
CCTGGAGA[ACGT]{0.50}TCTCCAGG	1	1	0.000	0.036	739.786
ATGACTGC[ACGT]{0.50}CGAGTCAT	1	1	0.000	0.036	739.786
AGGGGGA[ACGT]{0.50}JAGGGGAC	1	1	0.000	0.036	739.786
TCACAACA[ACGT]{0.50}TCACAACA	1	2	0.000	0.036	369.893
TAGCGTGG[ACGT]{0.50}CCACGCTA	1	2	0.000	0.036	369.893
GATCGATG[ACGT]{0.50}CATCGATC	1	2	0.000	0.036	369.893
GAGCTCT[ACGT]{0.50}GAGCTCTC	1	2	0.000	0.036	369.893
GAAGCAG[ACGT]{0.50}GCTGCTTC	1	2	0.000	0.036	369.893
CCCAACCG[ACGT]{0.50}CCCAACCG	1	2	0.000	0.036	369.893
CAACAACC[ACGT]{0.50}CAACAACC	1	2	0.000	0.036	369.893
AGGATTA[ACGT]{0.50}AGGATTA	1	2	0.000	0.036	369.893
AGACACAG[ACGT]{0.50}CTGTGTCT	1	2	0.000	0.036	369.893
TGTGCC[ACGT]{0.50}TGTGCC	1	3	0.000	0.036	246.595
GGGGCAC[ACGT]{0.50}GGGGCAC	1	3	0.000	0.036	246.595
GATCGATT[ACGT]{0.50}AATCGATC	1	3	0.000	0.036	246.595
TTTTGTCT[ACGT]{0.50}TTTTGTCT	1	4	0.000	0.036	184.946
TATGACT[ACGT]{0.50}AGTCATA	1	4	0.000	0.036	184.946
GCTAAAA[ACGT]{0.50}GCTAAAA	1	4	0.000	0.036	184.946
CCTGGAG[ACGT]{0.50}CTCCAGG	1	4	0.000	0.036	184.946
ACAGGGGA[ACGT]{0.50}TCCCTGT	1	4	0.000	0.036	184.946
GACACAG[ACGT]{0.50}CTGTGTC	1	5	0.000	0.036	147.957
CTAATCAT[ACGT]{0.50}CTAATCAT	1	5	0.000	0.036	147.957
ATAATATC[ACGT]{0.50}ATAATATC	1	5	0.000	0.036	147.957
TTCGTGG[ACGT]{0.50}TTCGTGG	1	6	0.000	0.036	123.298
CTTGGCT[ACGT]{0.50}CTTGGCT	1	6	0.000	0.036	123.298
CTGGAGA[ACGT]{0.50}TCTCCAG	1	6	0.000	0.036	123.298
AGCGTGG[ACGT]{0.50}CCACGCT	1	6	0.000	0.036	123.298
AGACAAA[ACGT]{0.50}AGACAAA	1	6	0.000	0.036	123.298
GTTGGCA[ACGT]{0.50}TGCCAAC	1	7	0.000	0.036	105.684
CITTCAC[ACGT]{0.50}GTGAAAG	1	7	0.000	0.036	105.684
CCTGTCT[ACGT]{0.50}CCCTGCT	1	7	0.000	0.036	105.684
CCACCTCG[ACGT]{0.50}CCACCTCG	1	7	0.000	0.036	105.684
TTTGATT[ACGT]{0.50}TTTGATT	1	8	0.000	0.036	92.473
TTCCATT[ACGT]{0.50}TTCCATT	1	8	0.000	0.036	92.473
TATGGTGA[ACGT]{0.50}CACCATA	1	8	0.000	0.036	92.473
GAAAGCA[ACGT]{0.50}GAAAGCA	1	8	0.000	0.036	92.473
CTGAAG[ACGT]{0.50}CTTTACG	1	8	0.000	0.036	92.473
CGGGCAC[ACGT]{0.50}CGGGCAC	1	8	0.000	0.036	92.473
CGCCTGT[ACGT]{0.50}CGCCTGT	1	8	0.000	0.036	92.473
AATTTGG[ACGT]{0.50}AATTTGG	1	8	0.000	0.036	92.473
TTCITTA[ACGT]{0.50}TTCITTA	1	9	0.000	0.036	82.198
GGCCACT[ACGT]{0.50}GGCCACT	1	9	0.000	0.036	82.198
GAGCTCT[ACGT]{0.50}GAGCTCT	1	9	0.000	0.036	82.198
CCACAAG[ACGT]{0.50}CTTGTGG	1	9	0.000	0.036	82.198
TCAATTA[ACGT]{0.50}TCAATTA	1	10	0.000	0.036	73.979
CCTCTGC[ACGT]{0.50}GCGAGGG	1	10	0.000	0.036	73.979
CCTCTGC[ACGT]{0.50}GCGAGGG	1	10	0.000	0.036	73.979
ATTGTG[ACGT]{0.50}ACACAAT	1	10	0.000	0.036	73.979
ATACACT[ACGT]{0.50}ATACACT	1	10	0.000	0.036	73.979
AGGGGCA[ACGT]{0.50}AGGGGCA	1	10	0.000	0.036	73.979
AAAGGGGA[ACGT]{0.50}TCCCTTT	1	10	0.000	0.036	73.979

12. Result (2)

Gene list 2:

MEME listed 6078 motifs, of which 4069 showed high lift value. There were 6 motifs showed partial hit to records of "PRHA binding sites" in ATCIS (Table 2). While TGTTC was not listed, similar motifs, TGTCTC and TGTCTT were found with lift=1.107 and 1.179, respectively.

Possible 18234 combined sequence pattern were searched on upstream sequence of listed genes, and 2914 of them were found. Table 4 shows 50 motifs of them showing best lift values.

Table 3: Result of single motif search for TUs labelled with "GO:9733" in Kome database.

Motif	Hit TU	Hit TU	Support	Confidence	Lift	ATCIS Description
	in target group	in the whole				
TATACACA	4	350	0.000	0.030	1.761	PRHA BS in PAL1
ACACAC	24	2871	0.001	0.179	1.288	PRHA BS in PAL1
TCATAC	19	2355	0.001	0.142	1.243	PRHA BS in PAL1
TACATT	28	3843	0.001	0.209	1.123	PRHA BS in PAL1
TCATACA	7	983	0.000	0.052	1.097	PRHA BS in PAL1
ATACAT	30	4335	0.001	0.224	1.066	PRHA BS in PAL1

Table 4: Top 50 motifs in lift values for TUs labelled with "GO:9733" in Kome database.

Motif	Hit TU	Hit TU	Support	Confidence	Lift
	in target group	in the whole			
TGCACAGTACCTT[0.50]ACTGTGGA	1	1	0.000	0.007	154.582
TCTTTCACACCTT[0.50]G	1	1	0.000	0.007	154.582
TCGAAACACCTT[0.50]GTTTCGA	1	1	0.000	0.007	154.582
TCAGTTTTACCTT[0.50]AAACTGA	1	1	0.000	0.007	154.582
TAAATGCAACCTT[0.50]TAAATGGA	1	1	0.000	0.007	154.582
GCCTCTGACCTT[0.50]GCTCTGGA	1	1	0.000	0.007	154.582
GCCTGGCGACCTT[0.50]GCGCCCGC	1	1	0.000	0.007	154.582
CCATAGAGACCTT[0.50]CCATAGAG	1	1	0.000	0.007	154.582
CTAAGCAACCTT[0.50]TCTTATG	1	1	0.000	0.007	154.582
CAAGGAGACCTT[0.50]GCTCTCTG	1	1	0.000	0.007	154.582
ATTTATGGACCTT[0.50]CCATAAAT	1	1	0.000	0.007	154.582
TTTGGTTACCTT[0.50]AACCTAAA	1	2	0.000	0.007	77.291
TCTAATGACCTT[0.50]TCAITAGA	1	2	0.000	0.007	77.291
TCAGATCTACCTT[0.50]TCAGATCT	1	2	0.000	0.007	77.291
TCACCTTACCTT[0.50]TCACCTAT	1	2	0.000	0.007	77.291
TATTGTCAACCTT[0.50]TGACAATA	1	2	0.000	0.007	77.291
TANGAACACCTT[0.50]TAGAAGAC	1	2	0.000	0.007	77.291
GTGTGGCCACCTT[0.50]GCGCCACAC	1	2	0.000	0.007	77.291
GTAGCATCACCTT[0.50]GTAGCATC	1	2	0.000	0.007	77.291
GSTAAKACCTT[0.50]GSTAAKAC	1	2	0.000	0.007	77.291
GGAGCCACACCTT[0.50]GGAGCCCA	1	2	0.000	0.007	77.291
GGCGGGTACCTT[0.50]GGCGGGT	1	2	0.000	0.007	77.291
GGCTCTGACCTT[0.50]GCGAGCGC	1	2	0.000	0.007	77.291
GCATTTGACCTT[0.50]GCATTTGC	1	2	0.000	0.007	77.291
GCAGCAACACCTT[0.50]GCAGCAAC	1	2	0.000	0.007	77.291
GAGCTAGACCTT[0.50]GAGCTAGA	1	2	0.000	0.007	77.291
GAAGGAGACCTT[0.50]GCTGCTTC	1	2	0.000	0.007	77.291
GAAGAGACCTT[0.50]TCTCTTTC	1	2	0.000	0.007	77.291
CTTCACATACCTT[0.50]CTTCACAT	1	2	0.000	0.007	77.291
CCATGCTTACCTT[0.50]AACCTGCG	1	2	0.000	0.007	77.291
CATTGTTACCTT[0.50]CATTGTT	1	2	0.000	0.007	77.291
CAGTAGACCTT[0.50]GCTACTG	1	2	0.000	0.007	77.291
CAATAGACCTT[0.50]CAATAGAC	1	2	0.000	0.007	77.291
CAACATCCACCTT[0.50]CAACATCC	1	2	0.000	0.007	77.291
ATATCTAGACCTT[0.50]ATCTAG	1	2	0.000	0.007	77.291
AGTCAATACCTT[0.50]AGTTCATA	1	2	0.000	0.007	77.291
AATTATGACCTT[0.50]CAATAAAT	1	2	0.000	0.007	77.291
TTTAGCTACCTT[0.50]TTTAGCTC	1	3	0.000	0.007	51.527
TGCTTGCACCTT[0.50]TGCCTTGC	1	3	0.000	0.007	51.527
TCTGGCGACCTT[0.50]TCTGGCGC	1	3	0.000	0.007	51.527
TCCTCTGACCTT[0.50]GCGAGGGA	1	3	0.000	0.007	51.527
TACTACACACCTT[0.50]TACTACAC	1	3	0.000	0.007	51.527
GCAGAAATACCTT[0.50]GCAGAAAT	1	3	0.000	0.007	51.527
GCTCTGACCTT[0.50]GCTCTG	1	3	0.000	0.007	51.527
GCTAGTGCACCTT[0.50]GCTAGTTG	1	3	0.000	0.007	51.527
GCCAKAAACCTT[0.50]GCCAKAAA	1	3	0.000	0.007	51.527
GAGCCCAACCTT[0.50]GAGCCCAA	1	3	0.000	0.007	51.527
GAGAAATACCTT[0.50]AATTTCTC	1	3	0.000	0.007	51.527
CTTCACCTACCTT[0.50]CTCACCT	1	3	0.000	0.007	51.527
CCCTGGGACCTT[0.50]CCGTGGCG	1	3	0.000	0.007	51.527
AGCGGGACCTT[0.50]AGCGGGCG	1	3	0.000	0.007	51.527
AGCTAGAACCTT[0.50]AGCTAGAA	1	3	0.000	0.007	51.527
AGGCTCGACCTT[0.50]AGGCTCGA	1	3	0.000	0.007	51.527
ACCCTGTACCTT[0.50]ACCCTGTGA	1	3	0.000	0.007	51.527
ACAGAGGACCTT[0.50]ACAGAGGG	1	3	0.000	0.007	51.527
ACAGAGAACCTT[0.50]ACAGAGAA	1	3	0.000	0.007	51.527
AATTCTCACCTT[0.50]GAGAAATT	1	3	0.000	0.007	51.527
AATCTATAACCTT[0.50]TATAGATT	1	3	0.000	0.007	51.527
AAATCTGACCTT[0.50]TCAAGATT	1	3	0.000	0.007	51.527

13. Discussion

In the presented test, TGTCTC element of 'AuxRE' was listed in the single motif search for Aux/IAA genes. The result is consistent with previous reports. This motif may act as composite motif as described in Ulmasov et al (1995), because positive relationship of AuxRE for genes was not shown in the combined motif analysis. Similar motifs such as TGTCGC and TGTTC were found to show high lift value. It is possible that these motifs play important role for auxin-responsive reaction.

After survey of papers and public databases as to found motifs, we realized that little is known about what kind of cis-elements concerned in regulation of auxin-inductive genes.

The approach we proposed here is useful to list up motifs of possible cis-elements rapidly. Using a Pentium computer of average power, it takes one day at the maximum to list up thousands of cis-element candidates with relating information, from gene lists of the same size of that used in the present examination.

Further refinement of the algorithm and parameters are planned to achieve more effective analysis. We are positive that the proposed mining tool has the potential to contribute much to accelerate the research of cis-element and gene expression regulatory network.