DEVELOPMENT OF A DATA MINING TOOL TO FIND CIS-ELEMENTS CONCERNING IN GENE EXPRESSION DETECTABLE BY MICROARRAY ANALYSIS

Doi K.[1], Nagata T.[1], Satoh K.[1], Suzuki K.[2], Iizumi S.[1], Kimura S.[1], Hosaka A.[1] and Kikuchi S.[1] 1：National Institute of Agrobiological Sciences, Japan. 2：Hitachi Software Engineering, Japan.

- 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. Such motifs are expected to play key roles in gene 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.
- This tool is developed under the “Generation Challenge Programme”, and planed to be opened publicly with www-interface in near future.

Introduction of the novel tool

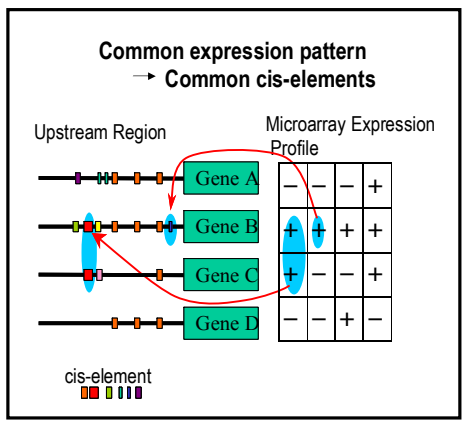
Motivation and Goal

NEEDS
List of Cis-elements relating to focused gene expression pattern

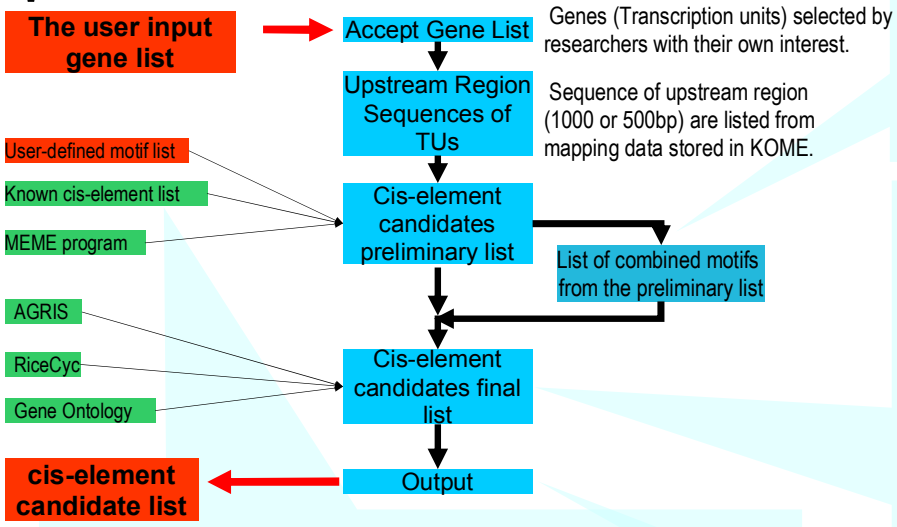
→

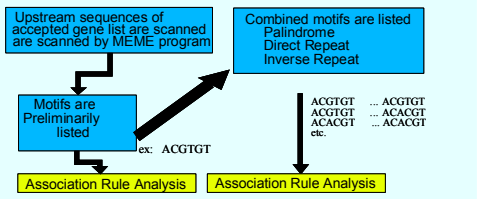
NEW TOOL
Combination of motif search and association rule analysis

Basic Concept
If the motifs specifically popular in the specified gene group, they should play specific roles on expression regulation of those genes.



Specification of the tool





Association Rule Analysis

Reliability of relationships(rules) can be quantitatively evaluated by indexes, such as Support, Confidence and Lift.

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.

Example:

		Y		Total
		Yes	No	
X	Yes	2	5	7
	No	0	3	3
Total		2	8	10

$$Support = \frac{2}{10} = 0.20$$

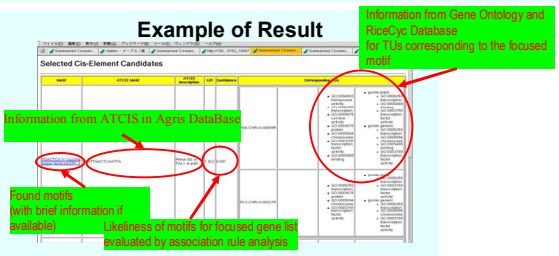
$$Confidence = \frac{2}{7} = 0.29$$

$$Lift = \frac{2/7}{7/10} = 0.41$$

High lift value suggests strong relationship between X and Y.

- **List of Known Cis-Element**
  - General information about cis-element have been collected based on experimental results of transcription factors as follows:
  - Binding to common sequences in many organisms-bHLH, bZIP >G-box, E-box.
  - Generally binding to common sequences in many organisms-Homeo domain, Myb >A-box, T-box, GGTTTAG Repeats
  - Binding in plant-MADS, zinc finger, AP2/EREBP > CArG boxes, GCC-box etc.
  - Binding only in animal- HSF, PcG, HMG etc.

Example of Result



Feasibility Test: A Case Study with Auxin Responsive Genes

Background

- Biologically important genes relating to plant growth, development, etc.
- A cis-element 'AuxRE‘ (=TGTCTC), known to regulate expression of auxin responsive genes, is a simple example of multiple cis-element interaction for gene expression regulatory.

Data Set

Aux/IAA genes of *Oryza sativa* stored in RiceTFDB (2.0) (http://ricetfdb.bio.uni-potsdam.de/v2.0/) :

- Aux/IAA genes themselves are auxin-inducible.
- Blastn search with these sequences to GenBank resulted 28 corresponding rice transcription units (TUs). → **Data set used to examine the tool.**

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 genes *2	Confi-dence	Lift	ATCIS Description *3
ACACAC	7	2873	0.25	1.802	PRHA BS in PAL1
ACATTA	10	3232	0.357	2.289	PRHA BS in PAL1
ACATTAT	4	1190	0.143	2.487	PRHA BS in PAL1
ACTCAA	4	2842	0.143	1.041	PRHA BS in PAL1
ATACAC	7	2446	0.25	2.117	PRHA BS in PAL1
ATACACAC	3	347	0.107	6.396	PRHA BS in PAL1
ATACATT	3	1260	0.107	1.761	PRHA BS in PAL1
CAATTA	6	3775	0.214	1.176	PRHA BS in PAL1
TACACA	6	3148	0.214	1.41	PRHA BS in PAL1
TACATT	7	3842	0.25	1.348	PRHA BS in PAL1
TACATTA	3	993	0.107	2.235	PRHA BS in PAL1
TATACA	10	3802	0.357	1.946	PRHA BS in PAL1
TATACACA	2	351	0.071	4.215	PRHA BS in PAL1
TGTCTC	4	2088	0.143	1.417	ARF1 binding site motif
TTATACAC	1	202	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
*4 The number of transcription units associating to the designated motif, of 28 'AUX/IAA' TUs.

Table 2: Cis-element candidates likely corresponding to AUX/IAA genes, selected from the list of known cis-element.

Motif	Category	Lift	TU number*4
([ACGT]GAA[ACGT]){3}	HSF.	4.244	3
TGACAGGT	Helix-turn-helix(HTH).	4.18	3
CCAC[AC]A[ACGT][AC][ACGT][CT][AC]	LIM finger.	2.586	11
● GG[ACGT]CCCAC	Helix-loop-helix factors(bHLH).	2.283	10
● GTGG[ACGT]CCC	Helix-loop-helix factors(bHLH).	2.192	6
● CAACA[ACGT]*CACCTG	RAV.	1.881	5
AATATATTT	Helix-turn-helix(HTH).	1.712	3
● TGTCTC	ARF.	1.521	8
TGACGTGG	NAC.	1.328	1
● CCA[ACGT]TG	LEAFY.	1.26	20
[CT]AAC[GT]G	Myb.	1.217	17
CACCC	RING finger.	1.216	19
AATAAA[CT]AAA	Helix-turn-helix(HTH).	1.154	1
CCAAT	Co-like.	1.086	20
CC([AT]){6}GG	MADS.	1.079	2
CGTG[TC]G	BZR(BES1).	1.05	9
[TA]AAAG	Dof.	1.039	27
● CA[ACGT][ACGT]TG	Helix-loop-helix_leucine zipper factors(bHLH-ZIP).	1.026	28
(T){4,6}	JUMONJI.	1.016	28
G[GT]T[AT]G[GT]T	Myb.	1.004	12
TAAT	JUMONJI.	1.001	27

MEME listed 5246 motifs from their upstream sequences (1000bps), of which 4579 TUs showed high lift value (>1.0). There were16 TUs showing partial match to the ATCIS records of “PRHA binding sites”. One of them matched to “ARF1 binding site”, of which sequence is same as that of “AuxRE”.

Table 2 shows that many kind of cis-elements associating of gene expression of auxin-inducible genes. Some of them have been suggested by previous studies to have the relationships to auxin response(● in Table 2). For example, RAV1 have been found in the promoter region of ABP gene encoding auxin-binding protein. Gene expression of LEAFY is affected by auxin gradient. ETT gene product, homologous to ARE-binding proteins ARF1 and IAA24, acts as transcription factor activating LFY gene. It is also suggested that bHLH families may have significant roles for hormone-induced response.

These results suggest the usefulness of our tool to survey possible combination of cis-elements involving gene expression regulation.

Many possible cis-elements have not been verified experimentally for their function for certain gene expression. Using the approach proposed here, researchers can rapidly list up motifs of possible cis-elements for further examination, to understand genetic mechanism of the biological phenomena that they are interest in.