THE DEVELOPMENT OF A NOVEL DATA MINING TOOL TO FIND *CIS*-ELEMENTS IN RICE GENE PROMOTER REGIONS

Koji Doi(1), Aeni Hosaka(1), Toshifumi Nagata(1), Kouji Satoh(1), Kohji Suzuki(2), Ramil Mauleon(3) , Michael Jonathan Mendoza(3), Richard Bruskiewich(3), and Shoshi Kikuchi(1)
(1) National Institute of Agrobiological Sciences (2) Hitachi Software Engineering, Japan.Co., Ltd. (3) International Rice Research Institute

Basic Feature of the Developed Tool

Information about over 35,000 full-length *Oryza sativa* cDNA and associated microarray gene expression data have enabled identification of conserved motifs in promoters of genes listed by microarray analysis. They are expected to act as *cis*-regulatory elements involved in key roles under the assayed experimental conditions.

We developed and have continued improvement of a novel tool to search *cis*-element candidates. Here we report the recent progress.

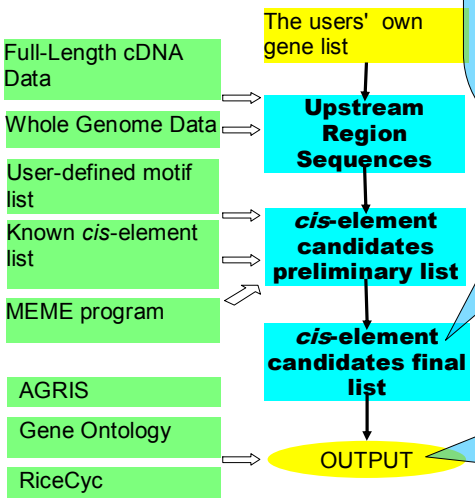
The Important Features of the Developed *cis*-Element Search Tool

- It is publicly opened as a web application software.
- No special programming skill is required.
- Users can define own gene set. (e.g. by cluster analysis after microarray analysis)
- Motif search (MEME) and data mining method (association rule analysis) are performed by a pipeline system implemented in the host computer in IRRI.

Basic Strategy for Search and Evaluation of *cis*-Element Candidates

If motifs overrepresented in the upstream region of the focused genes, they could play specific roles on expression regulation of the genes.

Flowchart of *cis*-Element Search



Evaluation to Get Final *cis*-Element Candidate List

Candidates showing the highest likelihood (specificity) are retained in the final cis-element candidate list.

→ Association Rule Analysis

X: The focused TU possess a cis-element candidate TGTTCT.
Y: The focused TU is corresponding to drought response.

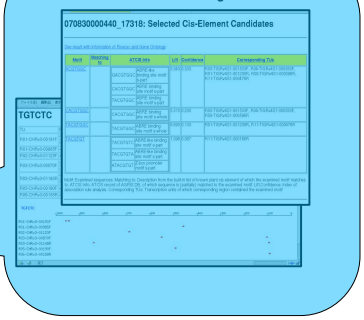
A Fictional Example:

	Y: Yes	Y: No	Total
X: Yes	8	176	184
X: No	1003	19322	20325
Total	1011	19498	20509

Support = 8 / 20509 = 0.00039
Confidence = 8 / 184 = 0.04348
Lift = 0.00039 / (184/20509)*(1011/20509) = 0.88199

High lift value suggests strong relationship between X and Y.
The developed tool defines default threshold lift value as 1.0, while users can change it.
Thus, as for example above, relationships between TGTTCT cis-element and drought stress is not strongly supported.

Screenshot of *cis*-Element Search Tool Entrance Page



Recent Upgrade of the Tool

Users can now select reference gene sets

Available or Preparing Reference Gene Sets:

- Whole set – including all genes in KOME database (default).
- Sets of genes including particular motifs.
- Sets of genes up-regulated in particular experimental condition (under construction).

We have confirmed that the contents of the reference have a great influence on result. Refinement of reference gene sets are still in progress.

Users can also specify *cis*-element candidate motifs

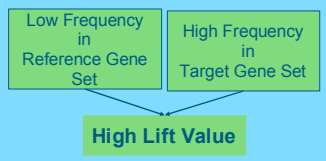
Perl-compatible regular expression powerfully support users to specify complex sequence patterns.

Example of regular expression for known *cis*-elements

Helix-turn-helix(HTH)	(CTAATTG){2,3}
BBR/BPC	((GA)+|(TC)+)
RAV	CAACA[ACGT]*CACCTG

Importance to Prepare Reference Data Set Alternatives

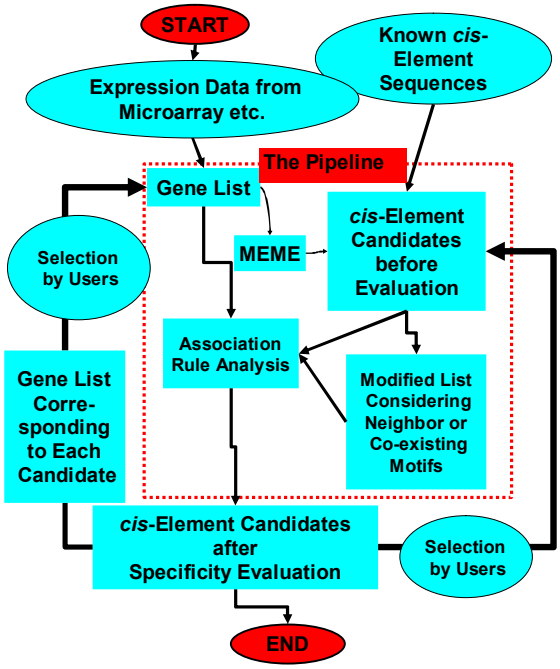
Lift value: The index of 'specificity' of the focused motif.



Typical situation in analysis

Any cis-element candidate motifs look unique and "specific" in the focused gene sets.
→ **High Lift Value, False Positive Result**

Advanced Data Mining Flowchart with Repeated Utilization of the *cis*-Element Search Tool



Improvement of the method to obtain more accurate result

Case Study for Feasibility Test of the Methods

Focused Topics:
Drought Stress Response – An important biological activity for survival of plants.
Abscisic Acid (ABA) Response – Closely related mechanisim of drought stress response.

There are some known *cis*-elements that play roles on these activities, while it is considered that many unknown *cis*-elements still remain.

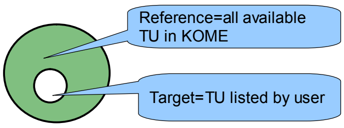
It is suitable for case study to evaluate feasibility of our method and developed tool.

We are considering new method to increase feasibility of the tool. Feasibility test is performed for the considered method.

Current Issue and Method under consideration

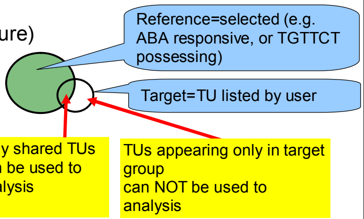
Default

Extraordinary large lift values tend to be obtained: Every candidate appears "unique"



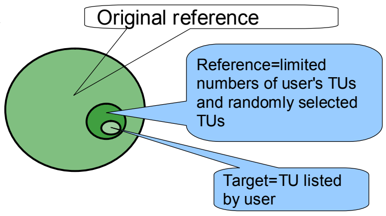
Switched Reference
(newly implemented feature)

Very few samples can be investigated: Result with Low reliability



New approach under consideration

Clear result is expected



Material and Method

Reference TU lists

- Two reference TU lists were prepared by literature survey.
  - TUs containing ABRE motif (ACGTG[GT]C) in 1000bp of upstream region. (804TU).
  - TUs containing ABA-responsive cis-elements motifs listed by literature survey in 1000bp of upstream region(11134TU)

User-defined TU lists

- Rabbani et al. (2003) showed a gene list (based on NCBI accession numbers) of some kinds of stress responsibility. Followings were from their work, that used as user-defined TU list for the feasibility test.
  - ABA responsive genes (18TU)
  - Drought responsive genes (24TU)

Analysis and evaluation of result

- All combination of reference and user-defined TU lists mentioned above were applied to analysis.
- Result was checked by simple negative control test.

Negative Control Test

To check whether the found "positive" relationship can be considered true, following simple test was employed in this study.

- Select TUs randomly from the reference list, so that the size of result list is twice as large as user-defined TU list.
- Specificity of known cis-element motifs for user-defined TU list was evaluated with selected reference prepared in the former step by developed tool.
- Perform *cis*-element search.
- For randomly-selected TUs, lift values of *cis*-element listed in the examination should be negative, while they should be positive for real user-defined TUs.

An example of negative control test for Leucine zipper factors(bZIP: CACGTG)

Length of searched region (bps)	Reference TU list	Lift Value: "user-defined" TU list: ABA	Drought	Negative Control Test using randomly selected TUs									
50	ABRE	12.00	16.00				2.00				2.00		
1000		2.25	1.92							1.33			1.33
50	ABA respon-sive	2.77	none										
1000		1.70	none	1.43			1.33	2.00	1.33	1.14	1.33	1.60	
50		none	3.00			2.00						2.00	
1000		none	1.59		1.67			1.43			1.40		

Result of feasibility test

Examined TUs			ABA		Drought		ABA		Drought	
Reference TU Set			ABRE				ABAresp			
Length of examined sequence			50	1000	50	1000	50	1000	50	1000
Previously known cis-elements	AP2	TACGGACAT								
	ARF	TGTCTC	6.00	3.43	16.00	3.20	2.77	1.19	3.00	
	ARID	TATTGAT						1.11		
	BRE	[GC][GC][GA]CGCC								
	BZR(BES1)	CGTG[TC]G	6.00	1.33	16.00		2.77		1.50	
	CCAATbox	CCAAT	4.00	1.20	2.67	1.33	1.39			
	Cx7Hx8Cx4C zinc finger	[AG]TC[AG][CT][ACGT][ACGT][ACGT][ACGT][ACGT]ACG						1.39		
	Cys2His2 zinc finger	GGGCGG		1.09		1.78		1.19		1.13
	Cys2His2 zinc finger/Ringer	CACCC		1.50		1.10			1.20	
	Dof	[TA]AAAG	2.00	1.06	2.00	1.09		1.04	1.50	
	DPE	[AG]G[AT][CT][GAC]					1.01			
	E2F/DP	([CG]GCGC[GC]){2}						2.77		1.50
	EREBP	AGCCGCC								2.00
	GATA-Factors	[AT]GATA[AG]		1.71		1.55				
	Helix-loop-helix factors(bHLH)	CA[ACGT][ACGT]TG	3.00		4.00		1.39			1.02
	Helix-loop-helix factors(bHLH)	GTGG[ACGT]CCC		4.00		2.29		2.77		1.50
	Helix-loop-helix factors(bHLH)	GG[ACGT]CCCAC								
	Helix-turn-helix(HTH)	AATATATTT						1.39		
	Helix-turn-helix(HTH)	TT[AT]T[AT]TT[AT]TT								
	Helix-turn-helix(HTH)	TGACAGGT								
	Helix-turn-helix(HTH)	AATAAA[CT]AAA								
	Helix-turn-helix(HTH)	AATATT		1.41		1.39				
	JUMONJI	TATT					1.11			
	JUMONJI	TAAT		1.06		1.09	1.39	1.03	1.29	
	JUMONJI	T{4,6}		1.03		1.02	1.66	1.03	1.29	1.04
	LEAFY	CCA[ACGT]TG		1.26		1.60				1.24
	Leucine zipper factors(bZIP)	TGACGTCA						2.77		3.00
	Leucine zipper factors(bZIP)	CACGTG	12.00	2.25	16.00	1.92	2.77	1.70	3.00	1.59
	Leucine zipper factors(bZIP)	T[TG][ACGT][ACGT]G[ACGT]AA[TG]				1.03				
	LIM finger	CCAC[AC]A[ACGT][AC][ACGT][CT][AC]					2.77	1.39	3.00	1.15
	MADS(CArG boxes)	CC[AT]{6}GG		3.00		2.67		1.39		2.25
	Myb	G[GT]T[AT]G[GT]T				1.07		1.07		
	RAV	CAACA[ACGT]*CACCTG		3.00		2.00		1.85		1.80
	TATAbox	TATA[AT]A		1.29		1.26	1.07		1.31	
	VOZ-9	GCGT[ACGT]{7}ACGC								
	WHIRLY/WRKY	GTCAAAA[AT]						1.85		1.50
	WRKY	(TT)?TGAC[CT]		1.24		1.50		1.09		
	YABBY	CCGAC						1.39		1.36

cis-elements showing positive correlation

lift values supported by negative control test

lift values not supported by negative control test

Lift values of known *cis*-element motifs for each examination are shown in the table. Values lower than 1.0, recognized as negative, are not shown.

Previous Related Studies

As far as our literature survey, following studies are especially notable to consider the feasibility of our method.

WRKY – WRKY TFs are known as factors involving various biological reactions such as disease resistance. Otherwise, Mare et al. (2004) reported that a WRKY TF (Hv-WRKY38) is induced by drought stress in *Hordeum vulgare*.

bZIP - Responsibility of "water deficit stress" in *Phaseolus* is reported (Rodriguez-Uribe and O'Connell, 2006).

JUMONJI - Senthil-Kumar (2007) pointed out a JUMONJI TF showed relative drought tolerant phenotype in *Nicotiana benthamiana*.

As for LIM finger, we have not found reports discussing the relationships to stress response definitely.

Discussion

Due to refinement on demand, now users can perform trial and error from various angles in using the *cis*-element search tool. The experimental version of the developed tool is still maintained to try refinement of data mining protocol.

The result of feasibility test is concordent with some previous reports. The number of detected cis-elements was limited, but the result suggested that plausible candidates can be obtained after slight trial and error.

We consider that the usefulness of the developed tool is ensured more due to update of the tool. We are going to reflect such attempt to the opened tool.

Future Work

Refinement of the reference gene sets

- Literature search of known cis-elements, and preliminary search of found motifs to select genes containing them.
- Collaboration with experimental researchers to create precise gene list responsive to particular biological phenomena.

Biological discovery

- The project will successfully end when suggestion of novel cis-elements are persuasively achieved.

Availability of the tool
http://hpc.irri.cgiar.org/tool/nias/ces

Any comments are greatly appreciated
Koji Doi (kdoi@affrc.go.jp)