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

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

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)(G,3)
BRR/BCPC ((GATA)(T))
RAV CAACA[ACGT][CACCTG]

Recent Upgrade of the Tool

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

START
Expression Data from Microarray etc.
Known cis-Element Sequences
Gene List
Selection by Users
MEME
cis-Element Candidates before Evaluation
Gene List Corresponding to Each Candidate
Association Rule Analysis
Modified List Considering Neighbor or Co-existing Motifs
 cis-Element Candidates after Specificity Evaluation
Selection by Users
END
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 mechanism 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

- Reference TU lists were prepared by literature survey.
  - TUs containing ABRE motif (ACGTTG/CTGC) in 100bp of upstream region. (BI04TU)
  - TUs containing ABA-responsive cis-elements motifs listed by literature survey in 1000bp of upstream region (11134TU)

User-defined TU lists

- Rabban 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 list was compared. Lift value 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 list 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 family (LZP, CACGTG)

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 (Hw-WRKY38) is induced by drought stress in Hordeum vulgare.

bZIP – Responsibility of "water deficit stress" in Phaseolus is reported (Rodriguez-Urbe and O'Connor, 2006).

JUMONJU – Senthilkumar (2007) pointed a JUMONJU TF showed relative drought tolerant phenotype in Nicotiana benthamiana.

As for LM 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 concordant 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.iri.cgiar.org/tool/nias/ces

Any comments are greatly appreciated

Koji Doi (kdoi@affrc.go.jp)