Development of a novel data mining tool to find cis-elements in rice gene promoter regions

Koji Doi¹, Aeni Hosaka¹, Toshifumi Nagata¹, Kouji Satoh¹, Kohji Suzuki², Ramil Mauleon³, Michael J Mendoza³, Richard Bruskiewich³ and Shoshi Kikuchi*¹

Address: ¹National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan, ²Hitachi Software Engineering Japan Co., Ltd., 6-81 Onoe-cho, Naka-ku, Yokohama 231-0015, Japan and ³International Rice Research Institute, DAPO 7777, Metro Manila, Philippines

Email: Koji Doi - kdoi@affrc.go.jp; Aeni Hosaka - aeni@nias.affrc.go.jp; Toshifumi Nagata - nagatat@nias.affrc.go.jp; Kouji Satoh - ksatoh@nias.affrc.go.jp; Kohji Suzuki - ksuzuki@hitachisoft.jp; Ramil Mauleon - rmauleon@yahoo.com; Michael J Mendoza - mvm4p@hotmail.com; Richard Bruskiewich - tbruskiewich@cgiar.org; Shoshi Kikuchi* - skikuchi@nias.affrc.go.jp

* Corresponding author

Abstract

Background: Information on more than 35,000 full-length Oryza sativa cDNAs, together with associated microarray gene expression data collected under various treatment conditions, has made it feasible to identify motifs that are conserved in gene promoters and may act as cis-regulatory elements with key roles under the various conditions.

Results: We have developed a novel tool that searches for cis-element candidates in the upstream, downstream, or coding regions of differentially regulated genes. The tool first lists cis-element candidates by motif searching based on the supposition that if there are cis-elements playing important roles in the regulation of a given set of genes, they will be statistically overrepresented and will be conserved. Then it evaluates the likelihood scores of the listed candidate motifs by association rule analysis. This strategy depends on the idea that motifs overrepresented in the promoter region could play specific roles in the regulation of expression of these genes. The tool is designed so that any biological researchers can use it easily at the publicly accessible Internet site http://hpc.irri.cgiar.org/tool/nias/ces. We evaluated the accuracy and utility of the tool by using a dataset of auxin-inducible genes that have well-studied cis-elements. The test showed the effectiveness of the tool in identifying significant relationships between cis-element candidates and related sets of genes.

Conclusion: The tool lists possible cis-element motifs corresponding to genes of interest, and it will contribute to the deeper understanding of gene regulatory mechanisms in plants.

Background

With the completion of rice genome sequencing by the International Rice Genome Sequencing Project [1], the Beijing Genomics Institute (BGI) [2], and Syngenta [3], many rice functional genomic resources have become available, including whole genome sequences from ssp. japonica ‘Nipponbare’ and ssp. indica line 93-11; a set of rice full-length cDNA clones and their complete and par-
tional end sequences [4,5], microarray gene expression systems based on full-length cDNA sequences, ESTs (Expressed Sequence Tag), MPSS (Massively Parallel Signature Sequencing), SAGE (Serial Analysis of Gene Expression), and predicted genes in the genome sequences; and many kinds of insertion mutants with Tos17, Ac-Ds, and T-DNAs [6]. As analytical technology progresses, the database continues to be upgraded and serves as a useful resource for studying mechanisms that regulate gene expression.

Cis-elements in the promoter regions of genes and trans-acting transcription factors are major biological features to be characterized if we are to achieve an understanding of the systems that regulate gene expression. Identification of candidate cis-elements corresponding to genes is now practicable through the use of available sequence and genome mapping information, combined with information about the responses of genes to specific experimental conditions; such responses have been elucidated by using gene expression profiles now publicly available.

Exhaustive sequence analysis by using available public databases can identify cis-element candidate motifs for further examination, but such approaches are not quite efficient. One confounding factor is that public databases are independently constructed and not generally optimized to facilitate integration of information from many sources with local experimental data. A more perplexing issue for experimental researchers who are not very familiar with bioinformatics techniques is the challenge of finding unknown but biologically notable relationships among genes, cis-elements, and experimental conditions from the huge number of possible combinations generated by large experimental data sets.

To resolve some of these issues, we developed a novel data mining tool to identify cis-elements in the rice genome. It performs the complex bioinformatics analysis mentioned above, then lists cis-element candidates for genes. The genes can be grouped by similarity of expression profiles and other criteria for assessment by researchers, then the tool annotates them with related public database information.

Similar tools have been developed previously. Helden released RSAT, which includes a program that can detect over-represented motifs in upstream regions of co-regulated genes [7]. Holt et al. established CoReg, which links the hierarchical clustering of co-expressed gene sets with frequency tables of promoter elements [8]. Zhao et al. established TRED, which integrates a database and a system for predicting cis- and trans-elements in mammals [9]. Galuschka et al. developed AthaMAP, which includes a program for comparative analysis of cis-elements in sets of co-transcribed genes of Arabidopsis thaliana [10].

Our tool is distinguished by several points: (i) It focuses on the rice genome, being based on full-length cDNAs, and is designed to pick up cis-element candidates associated with genes that users designate. (ii) It evaluates the likelihood score of cis-element candidates by comparing frequency counts in the user-selected gene set and a reference gene set. (iii) It can evaluate previously known cis-element sequences as well as user-specified sequences prepared by other analysis tools, and it can examine several cis-elements together.

The tool carries out both ab initio motif searches of promoter sequences and searches against known plant cis-elements, then performs a likelihood analysis of identified cis-elements on the basis of their presence in a significant proportion of the promoters of a given set of genes. This evaluation is achieved by an association rule analysis.

Here, we present technical details of the tool and demonstrate the practical assessment of its utility with a biologically relevant sample data set.

Implementation
The tool, called Rice Cis-Element Searcher (RiCES), consists of a cis-element searching pipeline, controlled via a Web-based user interface. Fig. 1 summarizes the procedure. The pipeline first reads a list of gene identifiers from the user, which it uses to retrieve the promoter sequences corresponding to the listed genes. Then a preliminary list of cis-element candidates is built by aligning information from the built-in list of plausible motifs, or by ab initio motif searching of the sequence data. Association rule analysis is carried out and reported to support the candidacy of the resulting cis-element list.

Gene list
RiCES assumes that a user has already identified genes of interest from experimental analysis (e.g. clusters of coordinately regulated genes). The list of identifiers is input into a Web-based data entry form. RiCES recognizes GenBank accession numbers, identifiers of transcription units (TUs) as defined in the TIGR pseudomolecular assemblies [11], and several other major gene identification systems. Using the list, it retrieves the set of associated upstream, downstream, or coding region sequences flanking the specified genes from available genomic sequence data.

Preliminary cis-element candidate list
The second step of the analysis is the compilation of a list of motifs as candidate cis-elements. At present RiCES supports two methods to achieve this.

Page 2 of 10
(page number not for citation purposes)
The first method depends on *ab initio* motif searching based on the supposition that if there are *cis*-elements playing important roles in the regulation of a given set of genes, they will be statistically overrepresented in the associated promoter sequences as conserved motifs that can be identified by using a suitable motif search program. There are several programs implementing several algorithms. We have chosen to use MEME, which is a publicly available motif discovery program [12] supporting an expectation maximization algorithm. In our analysis algorithm, MEME is invoked to identify motifs 6 to 8 bp long that look highly conserved among promoter sequences of the selected genes. Users can modify some of the search parameters of the MEME program via the Web form.

The second method relies on the hypothesis that common, known *cis*-elements play important roles under the experimental conditions that gave rise to the list of genes specified by the user. Therefore, RiCES searches for matches to a pre-compiled list of known *cis*-elements.

Several databases of plant *cis*-elements are publicly available. PLACE [13] is one of the most popular databases of known *cis*-elements in plant genomes. AtcisDB, a part of AGRIS [14], includes information on *cis*-elements involved in gene regulation in *Arabidopsis thaliana*.

Although these databases are extremely useful resources, it is not straightforward to cross-link information from them directly to the researcher’s own data. Current databases are not exhaustive enough to distinguish ‘core’ motifs, which decide the function of *cis*-elements, from co-existing sequences in neighboring regions. As a result, many *cis*-element sequence data in these databases include superficial core motifs for which no evidence of functionality has been obtained. The use of such data prohibits effective informatic analysis.

We compiled a novel database of known *cis*-elements and incorporated it into RiCES [See Additional file 1]. The *cis*-elements are collected from reports of experiments such as

---

**Figure 1**

**Features of RiCES.**
gel shift assays and footprint analyses, categorized by transcription factor, and documented with respect to known activity in the plant genome. Some cis-elements known only in organisms other than plants are also listed, in consideration of their possible, albeit unknown, roles in plants. The database includes four types of cis-elements: (1) G-box and E-box, which bind to common sequences such as bHLH or bZIP in many organisms; (2) A-box, T-box, and GTTTTAG repeats, which bind to common sequences in many organisms, such as homeodomain and Myb; (3) CArG boxes and GCC-box, which bind to plant MADS, zinc finger, and AP2/EREBP elements; and (4) other cis-elements, binding only in animals, such as HSF, PcG, and HMG.

**Association rule analysis**

The third step of the analysis is the likelihood evaluation of the cis-element candidates by association rule analysis, which is a data mining method designed to discover significant relationships between pairs of characteristics observed in data sets. Candidates showing the highest likelihood (specificity) are retained in the final cis-element candidate list.

Association rule analysis has been applied to mechanisms that regulate gene expression [e.g. [15,16]]. We used it to find relationships between identified cis-elements and gene expression profiles. The strategy depends on the idea that motifs overrepresented in the promoter region of the genes of interest could play specific roles in regulation of the expression of those genes.

Implied cause-and-effect relationships documented as 'rules' are evaluated by using several well-known indices of likelihood, including support, confidence, and lift [15]. On the basis of sample data sets, the lift index appeared to best discriminate significant relationships between experimental conditions and cis-element candidates.

In a rule described as

> the presence of motif X in a gene implies that the gene is a member of group Y,

lift is the ratio of the posterior probability (the probability that the gene is in group Y if it possess motif X) to the prior probability (the probability of X possession, irrespective of the membership of Y). When lift > 1.0, the coexistence of X and Y is not a random occurrence, but suggests some causal relationship between them. If lift < 1.0, it is not considered probabilistically significant. Consequently, we set the default threshold of lift to 1.0, and the cis-element candidates are included in the final candidate list only if their lift value is higher than this threshold.

RICES also evaluates pairwise combinations of motifs in the preliminary candidate list (upper right-hand box in Fig. 1), in consideration of possible protein-protein interactions of multiple transcription elements binding cis-elements, as illustrated by experimental evidence [17,18].

**Validation**

To test whether or not the output of RiCES was meaningful, we validated it with a list of auxin-inducible genes with known characteristics, compiled from RiceTFDB 2.0 [21]. First, Aux/IAA genes stored in RiceTFDB were applied as queries in a BLASTN search [22] of GenBank, returning a list containing 28 rice TUs [See Additional file 2]. These genes were fed into the pipeline. When the MEME program was called, the length of target motifs was set to 6, 7, or 8 bases, the number of occurrences of each motif was set to 7, 14, or 21, and the search algorithm was set to 'zoops' to check zero or one occurrence per sequence. The outputs of each option setting were merged but not otherwise filtered.

**Results and Discussion**

Many Aux/IAA genes are auxin-inducible [23] and contain the TGTCTC element [24]. This element is commonly found in the upstream region of auxin-responsive genes. Thus, the detection of all instances of the motif by the pipeline could serve as a validation of the pipeline algorithm. The auxin-responsive element (AuxRE) containing the TGTCTC motif in some cases requires another proximal AuxRE for biological activity [17,25]. In other contexts, AuxRE functions only when it occurs with its palindromic components separated by 7 or 8 nucleotides [26].

In our validation test, MEME listed 7514 motifs in total from 1000 bp of the upstream sequences [See Additional file 3], of which 4128 showed a high lift value (>1.0) [See Additional file 4]. A search of AtcisDB for these motifs returned 4 showing a partial match to the record of ‘PRHA binding sites’ (Table 1), which is derived from the report of Plesch et al. [27], describing auxin-induced expression of the Arabidopsis prha homeobox gene. Another 4 motifs
Table 1: Cis-element candidate motifs from Aux/IAA genes and suggested to be auxin-induction related according to ATCIS.

| Motif      | Hit TU in target group | Hit TU in whole | Lift     | ATCIS Description |
|------------|------------------------|-----------------|----------|-------------------|
| ACACAC     | 10                     | 6056            | 1.353    | PRHA BS in PAL1   |
| ATACACA    | 5                      | 2124            | 1.929    | PRHA BS in PAL1   |
| ATACACATAC | 3                      | 739             | 3.326    | PRHA BS in PAL1   |
| TACACAC    | 4                      | 1786            | 1.835    | PRHA BS in PAL1   |
| CATGTCTCTC | 1                      | 303             | 2.704    | -                 |
| GTGTCTCTC  | 1                      | 722             | 1.135    | -                 |
| TGTCTCCGT  | 1                      | 178             | 4.603    | -                 |
| TGTCTCTGT  | 2                      | 263             | 6.231    | -                 |

*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 22943 TUs stored in KOME database.
*3 PRHA = Developmental and auxin-induced expression of the Arabidopsis prha homeobox gene.

Table 2: Cis-element candidates selected from the pre-compiled list, likely corresponding to Aux/IAA genes.

| Motif      | Transcription Factor Family | Hit TU in target group | Hit TU in whole | Lift |
|------------|-----------------------------|------------------------|-----------------|------|
| ([ACGT]GAA ([ACGT])3) | HSF | 4 | 512 | 6.40 |
| TGACAGGT   | Helix-turn-helix(HTH)       | 3                     | 527             | 4.66 |
| CCACA      | LIM finger                 | 9                     | 3013            | 2.45 |
| GG ([ACGT]CCCAC | Helix-loop-helix factors(bHLH) | 10 | 3601 | 2.28 |
| GTGG ([ACGT]CCC | Helix-loop-helix factors(bHLH) | 6 | 2189 | 2.25 |
| CAAAC       | RAV                        | 8                     | 3039            | 2.16 |
| [T][G][A]TA[CT]CT | EIL                  | 3                     | 1405            | 1.75 |
| TGTCTCG    | ARF                        | 7                     | 3825            | 1.50 |
| TGACGTGG    | NAC                        | 1                     | 627             | 1.31 |
| CCA ([ACGT]TG | LEAFY                   | 19                    | 12084           | 1.29 |
| CACCC       | Cys2His2 zinc finger;RING finger | 19 | 12165 | 1.28 |
| CC ([AT][6]GG | MADS(CArG boxes)         | 2                     | 1392            | 1.18 |
| AATAAA      | Helix-turn-helix(HTH)      | 1                     | 715             | 1.15 |
| CGTG [TC]G | BZR(BES1)                 | 9                     | 6544            | 1.13 |
| [GC][GC][GA]CGCC | BRE       | 10                    | 7543            | 1.09 |
| AGCGGCC      | EREBP                     | 2                     | 1523            | 1.08 |
| CCAAT       | CCAATbox;Co-like          | 19                    | 14497           | 1.07 |
| TATA [AT]A | TATAbox                   | 22                    | 16849           | 1.07 |
| [TA]AAAG    | Dof                        | 27                    | 21329           | 1.04 |
| CA ([ACGT] [ACGT]TG | Helix-loop-helix factors(bHLH);Helix-loop-helix_leucine zipper factors(bHLH-ZIP) | 28 | 22405 | 1.02 |
| T(4,6)      | JUMONJII                  | 28                    | 22699           | 1.01 |
| [CT][CT]A [ACGT] [TA][CT]CT | Inr | 28 | 22899 | 1.00 |
| (GA)(2,2)(TC)(2,2) | BBR/BPC | 28 | 22911 | 1.00 |

*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 22943 TUs stored in KOME database.

contained the TGTCTC element. The result was consistent with previous work, as TGTCTC was listed as a candidate in the single motif search of Aux/IAA genes.

Table 2 shows the result of the validation test with a pre-compiled cis-element list generated by the test gene list. The analysis returned 22 cis-element candidates with lift > 1.0 [See Additional file 5 and 6]. Some of these candidates were suggested by previous studies to have some kind of relationship to auxin response. For example, RAV1 was found in the promoter region of ABP, which encodes an auxin-binding protein [28]. Expression of LEAFY (LFY) is affected by the auxin gradient in Arabidopsis [29]. ETT is another auxin response factor [30], and LFY and ETT expression are closely correlated [18,31].

The position of a cis-element is important information to consider in relation to the function of the cis-element. For biological activity to occur, the distance of some cis-elements from the coding region or other collaborating elements is constrained. To this end, RiCES highlights the distribution of cis-element candidates. It provides tables...
of identified cis-element motifs and graphical motif maps to help researchers grasp positional relationships among the candidate elements.

The positions of the listed elements, some of which include TGTCTC, varied among upstream regions of genes (Fig. 2), and it was hard to detect any skewed distribution of motifs. Goda et al. [32] studied the distribution of TGTCTC motifs in the genome of A. thaliana, and pointed out that 25% of investigated genes had TGTCTC motifs in the upstream region within 1000 bp of the start codon, and 14% within 500 bps. Our results do not seem in conflict of theirs.

TGTCTC motifs are scattered over wide regions of many plant species (Table 3). It is possible that the variety of the roles of genes reflects the variety of mechanisms regulating gene expression and positions of cis-elements, even if the genes in question can be classified as ‘auxin-responsive genes’ in a larger sense.

A major research concern is how to pick up cis-element candidates worthy of further experimentation. Computational and manual selection of cis-element candidates should play complementary roles to resolve this issue. It should be emphasized that cis-element candidates listed by RiCES are rated according to the likelihood provided by association rule analysis. On the other hand, researchers can check the significance of candidates in detail by using related information derived from several databases. The supported databases include AGRIS, Gene Ontology,
and RiceCyc, as well as the map information described above.

Fig. 3 is an example of the output for the TGTCTC motif. The outputs are not only easily accessible in a Web browser, but are also usable in further statistical or bioinformatics analysis, as they are also provided in XML format (Fig. 3A), which is a tagged plain-text format compatible with various computer programs.

In some cases, the results of the analysis from the pre-compiled list of elements will be easily comparable with prior knowledge. In other cases involving solely ab initio evidence from MEME, the results of motif searches should be interpreted carefully, because the result will change considerably in accordance with the options selected. An appropriate set of motif search options should be determined each time, by trial and error. However, as described above, a motif search can find cis-element candidates of which the sequences do not exactly match those of known cis-elements.

Although RiCIES is focused on the role of cis-elements in *Oryza sativa* ssp. *japonica*, the methodology can be applied easily to studies of other plant species, or of other genome sequence motifs involving gene expression regulation, such as motifs in coding regions of genes or downstream of the gene sequence. Such work can be made possible by replacing the reference data set containing whole genes of rice with other data sets.

### Conclusion

We presented here a newly developed tool to search for cis-element candidates in a list of genes. A case study showed the applicability of the tool. The tool is easy to use and publicly available. We expect that its use will deepen understanding of the mechanisms that regulate gene expression in plants.

### Availability and requirements

RiCIES is accessible at http://hpc.irri.cgiar.org/tool/nias/ces by any JavaScript-capable browsers.

**Project Name:** Generation Challenge Programme Sub-programme 4

**Project Home Page:** http://www.generationcp.org/subprogramme4.php

**Operating system(s):** Platform independent

**Other requirements:** None

**Programming language:** Perl

### Table 3: Representative plant genes possessing TGTCTC element in corresponding upstream region.

| Gene (domain) | Position | Remarks | References |
|---------------|----------|---------|------------|
| GH3 (D4)      | -130~~-125 | The auxin-responsive soybean GH3 gene. Domain D4 and D1. | [17, 33] |
| GH3 (D1)      | -176~~-171 | | |
| OsBLE3        | -434~~-429 | Brassinolide-enhanced gene involved in cell elongation in rice through dual regulation by BL and IAA. | [34] |
| GhMyb7        | -75~~-70  | A cotton R2R3-MYB gene. The transcript level is increased by auxin in fiber cells in an in vitro ovule culture system. | [35] |
| PsPK2         | -1695~~-1690 | PINOID-like gene from *Pisum sativum*. Auxin and gibberellin positively regulate its expression. | [36] |
| 14-3-3        | -625~~-620,-531~~-526 | Promoter of the gene of 14-3-3 proteins, participating in cell cycle control, was investigated in *Solanum tuberosum*. | [37] |
| LCA1          | -1430~~-1425 | Ca2+-ATPase gene of *Lycopersicum esculentum* induced by ABA and IAA. | [38] |
| CMe-ACS2      | -106~~-101  | ACS (auxin-responsive 1-aminocyclopropane-1-carboxylate synthase gene) of Melon (*Cucumis melo*). | [39] |
| OsRRAA1       | -150~~-145  | OsRRAA1 (*Oryza sativa* Root Architecture Associated 1) functions in the development of rice root system. | [40] |
| PsNin         | -364~~-359  | Genes function in early stages of root nodule formation in *Pisum sativum* (PsNin) or in *Lotus japonicus* (LjNin). | [41] |
| LiNin         | -365~~-360  | | |
| SAUR          | -134~~-129  | SAUR (Small Auxin-Up RNA) gene of *Glycine max*. | [42] |
| CEVI1         | -959~~-954,-119~~-114 | Defense-related CEVI1 gene is found from tomato (*Lycopersicon esculentum*). | [43] |
| EXPA1         | -2090~~-2085 | TSI (tropic stimulus-induced) genes observed in *Brassica oleracea*. | [44] |
| SKS1          | -1204~~-1199 | | |
| SAUR50        | -101~~-96   | | |
| GH3.5         | -86~~-81,-585~~-580 | | |
| AAP8          | -918~~-913  | *Arabidopsis* amino acid transporters (AAPs); AAP8 is probably responsible for import of organic nitrogen into developing seeds. | [45] |

*) Numbers are equivalent to those shown in the main text.
Authors' contributions

KD designed the algorithm, did all the programming, and performed the feasibility test of the tool. AH helped to prepare test data sets and the literature search. TN supplied the inner database of known cis-elements to which the tool refers. KSa and KSu prepared the reference data. RM, MJM, and RB made many technical suggestions on the implementation and set up the host computer. RB also corrected the English of this manuscript. SK conceived the

Figure 3

Snapshots of representative outputs of RiCES. A: List of cis-element candidate motifs including related information. B: Mapping image of cis-element candidate motifs.

License: Freely available for use

Any restrictions to use by non-academics: None

Authors' contributions

KD designed the algorithm, did all the programming, and performed the feasibility test of the tool. AH helped to prepare test data sets and the literature search. TN supplied the inner database of known cis-elements to which the tool refers. KSa and KSu prepared the reference data. RM, MJM, and RB made many technical suggestions on the implementation and set up the host computer. RB also corrected the English of this manuscript. SK conceived the
study and participated in its design and coordination. All authors read and approved the manuscript.

Additional material

Additional file 1
Known plant cis-elements listed for analysis by RiCES. See text for further details. Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-8-20-S1.xls]

Additional file 2
Transcription units (TUs) used in the feasibility test. Auxin-inducible genes were picked up from RiceTFDB 2.0 (1st column). Corresponding full-length cDNAs were designated by BLASTN (2nd column) and translated to TUs defined in Pseudomolecule ver. 4 (3rd column). Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-8-20-S2.csv]

Additional file 3
Preliminary list of cis-element candidates listed by MEME analysis for TUs shown in Supplementary Table S2. See text for further details. Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-8-20-S3.txt]

Additional file 4
Result of association rule analysis of cis-element candidates listed by MEME. 1st column: examined sequence. 2nd column: number of TUs possessing the designated motif within 28 TUs of the target gene list. 3rd column: number of TU possessing the designated motif within 22 943 TUs stored in KOME database. 4th column: lift value. Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-8-20-S4.csv]

Additional file 5
Result of sequence search for motifs shown in Supplementary Table S1 in 22 943 TUs stored in KOME database. 1st column: examined TUs. 2nd column: motifs found in upstream region of TU. Other columns: position of motifs within the upstream region of each TU. Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-8-20-S5.csv]

Additional file 6
Result of association rule analysis after sequence search shown in Supplementary Table S6. 1st column: examined sequence. 2nd column: number of TUs possessing the designated motif within 28 TUs of the target gene list. 3rd column: number of TU possessing the designated motif within 22 943 TUs stored in KOME database. 4th column: lift value. Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-8-20-S6.csv]

Acknowledgements
This work was supported by a grant from the Generation Challenge Programme SP4 2005–32 project.

References
1. IRGSP Project: The map-based sequence of the rice genome. Nature 2005, 436:793-800.
2. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Xu J, Zhang K, Zheng X, Dong J, Zenni R, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Gao W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H: A draft sequence for the analysis of the rice genome (Oryza sativa L. ssp. indica). Science 2002, 296:79-92.
3. Goff SA, Riche D, Lan T, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Karrari F, Lange BM, Mougous J, Rong P, Suh J, Miguel T, Paszkowski U, Zhang S, Colomber S, Sun W, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Audey N, Rubano T, Turonen M, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S: A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 2002, 296:92-100.
4. Rice Full-Length cDNA Consortium: Collection, mapping, and annotation of over 20,000 cDNA clones from japonica rice. Science 2003, 301:376-379.
5. Satoh K, Doi K, Nagata T, Kishimoto N, Suzuki K, Otomo Y, Kawai J, Murakami K, Carninci P, Hayashizaki Y, Kikuchi S, Katagiri F, Zhang S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J: The map-based sequence of the rice genome (Oryza sativa L. ssp. japonica). Nature 2002, 418:72-76.
6. Hirochika H, Guiderdoni E, An G, Haing Y, Eun MY, Han C, Upadhya N, Ramachandran S, Zhang Q, Pereira A, Sudarsenan V, Leung H: Rice mutant resources for gene discovery. Plant Mol Biol 2004, 54:325-334.
7. van Helden J: Regulatory sequence analysis tools. Nucleic Acids Res 2003, 31:3593-3596.
8. Holt KE, Millar AH, Whelan J: ModuleFinder and CoReg: alternative tools for linking gene expression modules with promoter sequences motifs to uncover gene regulation mechanisms in plants. Plant Methods 2005, 1:8.
9. Zhao F, Xuan Z, Liu L, Zhang MQ: TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. Nucleic Acids Res 2005, 33:D1235.
10. Galuschka C, Schindler M, Bulow L, Mehli R: AthaMap web tools for the analysis and identification of co-regulated genes. Nucleic Acids Res 2007, 35:D857-D862.
11. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis H, B. Wortman J, Buell CR: The TIGR Rice Genome Annotation Resource: Improvements and new features. Nucleic Acids Res 2007, 35:D883-D887.
12. Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 1994, 2:28-36.
13. Higo K, Ugawa Y, Iwamoto M, Kobayashi T: Plant cis-acting regulatory DNA elements (PLACE) database. Nucleic Acids Res 1999, 27:297-300.
14. Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurzt M, Grotevold E: AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics 2003, 4:225.
15. Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, Pascual-Montano A: Integrated analysis of gene expression by association rules discovery. BMC Bioinformatics 2006, 7:54.

16. Conklin DJ, Jonassen I, Aasland R, Taylor VR: Association of nucleotide patterns with gene function classes: application to human 3' untranslatable sequences. Bioinformatics 2002, 18:182-189.

17. Plesch G, Stoermann K, Torres JT, Walden R, Somssich IE: RiceTFDB - Integrated analysis of gene expression by association rules discovery. BMC Bioinformatics 2006, 7:54.

18. Ishiki Y, Oda A, Yaegashi Y, Orihara Y, Arai T, Hirabayashi T, Nakagawa H, Sato T: Cloning of an auxin-responsive 1-aminocyclopropane-1-carboxylate synthase gene (CMe-ACS2) from melon and the expression of ACS genes in etiolated melon seedlings and melon fruits. Plant Sci 2000, 159:173-181.

19. Ge L, Chen H, Jiang JF, Zhao Y, Xu ML, Xu YY, Tan KH, Xu ZH, Chong K: Overexpression of OsRRA1 causes pleiotropic phenotypes in transgenic rice plants, including altered leaf, flower, and root development and root response to gravity. Plant Physiol 2004, 135:1502-1513.

20. Bai F, Watson JC, Walling J, Weeden N, Santner AA, DeMason DA: A novel transcription factor involved in plant defense endowed with protein phosphatase activity. EMBO J 2003, 22:3376-3384.

21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene Ontology: tool for the unification of biology. Nat Genet 2000, 25:25-29.

22. Aziz S, Bozic K, Moldenhauer G, Lohaus A, Pfeifer M, Winter S: The rice genome: knowledge and tools. Arabidopsis Book 2005, 3:0012.

23. Shimada Y, Yoshida S, Kiyosawa K: A comprehensive comparison of auxin-regulated and brassinosteroid-regulated genes in Arabidopsis. Plant Physiol 2004, 134:1555-1573.

24. Tiwari SB, Wang XJ, Hagen G, Guilfoyle TJ: ARF1, a transcription factor that binds to auxin response elements. Science 1997, 276:1865-1868.

25. Koga Y, Ohnishi Y, Kato N, Tada M, Takahashi T, Obuchi H, Kato T, Nishimura Y, Kato K, Ohnishi Y: A G-box-binding protein from soybean binds to the E1 auxin-response element in the soybean GBH promoter and contains a proline-rich repression domain. Plant Physiol 1997, 115:397-407.

26. Ulmasov T, Hagen G, Guilfoyle TJ: A transcription factor that binds to auxin response elements. Science 1997, 276:1865-1868.

27. Asahina M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene Ontology: tool for the unification of biology. Nat Genet 2000, 25:25-29.

28. Kagaya Y, Ohnishi Y, Kato N, Tada M, Takahashi T, Obuchi H, Kato T, Nishimura Y, Kato K, Ohnishi Y: A G-box-binding protein from soybean binds to the E1 auxin-response element in the soybean GBH promoter and contains a proline-rich repression domain. Plant Physiol 1997, 115:397-407.

29. Ezawa A, Hasezumi M, Iwashita H, Hasezumi K, Iwashita A: Association of nucleotide patterns with gene function classes: application to human 3' untranslatable sequences. Bioinformatics 2002, 18:182-189.

30. Single J, Noamani A, Oda A, Yaegashi Y, Orihara Y, Arai T, Hirabayashi T, Nakagawa H, Sato T: Cloning of an auxin-responsive 1-aminocyclopropane-1-carboxylate synthase gene (CMe-ACS2) from melon and the expression of ACS genes in etiolated melon seedlings and melon fruits. Plant Sci 2000, 159:173-181.

31. Ge L, Chen H, Jiang JF, Zhao Y, Xu ML, Xu YY, Tan KH, Xu ZH, Chong K: Overexpression of OsRRA1 causes pleiotropic phenotypes in transgenic rice plants, including altered leaf, flower, and root development and root response to gravity. Plant Physiol 2004, 135:1502-1513.

32. Bai F, Watson JC, Walling J, Weeden N, Santner AA, DeMason DA: A novel transcription factor involved in plant defense endowed with protein phosphatase activity. EMBO J 2003, 22:3376-3384.

33. Esmon CA, Tinsley AG, Ljung K, Sandberg G, Hearne LB, Liscum E: A gradient of auxin and auxin-dependent transcription precedes tropic growth responses. Proc Natl Acad Sci USA 2006, 103:236-241.

34. Okumoto S, Schmidt R, Tegeder M, Fischer WN, Rentsch D, Frommer WB, Koch W: High affinity amino acid transporters specifically expressed in xylem parenchyma and developing seeds of Arabidopsis. J Biol Chem 2002, 277:45338-45346.

Submit your manuscript here: http://www.biomedcentral.com/info/publishing_adv.asp

BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime.*
Sir Paul Nurse, Cancer Research UK

Your research papers will be:
- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Page 10 of 10

(page number not for citation purposes)