Special Issue

Database Note

A gene co-expression database for understanding biological processes in soybean

Yoshiyuki Ogata, Hideyuki Suzuki, Daisuke Shibata*

Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan
* E-mail: shibata@kazusa.or.jp   Tel: +81-438-52-3900   Fax: +81-438-52-3948

Abstract   The sobean genome has been decoded independently in the United States, China, South Korea, and Japan using varieties indigenous to those regions. Using an EST library of the variety indigenous to the United States, a DNA microarray technique was developed and gene expression data in microarray experiments have been accumulated in public databases such as the Gene Expression Omnibus database (GEO). Such data can be used for genome-wide bioinformatics approaches to co-expression analysis for predicting gene function. To extract co-expressed genes in soybean, we analyzed soybean gene expression data based on DNA microarray datasets obtained from GEO and developed a database for soybean gene co-expression analysis. Using the database, we illustrate the steps to retrieve co-expressed genes that may be involved in isoflavonoid biosynthesis and secondary cell wall biosynthesis. Our database is useful for extracting co-expressed genes, which may be involved in a particular biological process, and then for understanding the mechanisms of such processes in legumes.

Key words: Biological processes, co-expression analysis, gene expression, microarray, soybean.

Decoding of the genomes of *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000) and rice (International Rice Genome Sequencing Project 2005) have been completed and their genomic information has been utilized for genome-scale transcriptome analysis using DNA microarray technology. The soybean genome draft based on the cultivar "Williams 82" was released early in 2008 (http://phytozome.net/soybean) and a large collection of soybean full-length cDNAs based on the cultivar "Nourin No. 2" have been published by Umezawa et al. (2008). Soybean DNA microarray techniques have been developed on the basis of the "Williams 82" genome. Using microarray chips, recent studies (Panthee et al. 2007; Alvord et al. 2007; Cheng et al. 2008; Wong et al. 2009; Haerizadeh et al. 2009a; Haerizadeh et al. 2009b) have analyzed the soybean transcriptome and have deposited the analyzed data in public databases such as the Gene Expression Omnibus (GEO) database (http://ncbi.nlm.nih.gov/geo/), the ArrayExpress database (http://www.ebi.ac.uk/microarray-as/ae/), and the SoyXpress database (Cheng and Strömvik 2008; http://soyexpress.agrenv.mcgill.ca/). Such transcriptome analyses in soybeans have promoted gene-level understanding of various types of biological processes, such as secondary metabolism pathways. Eisen et al. (1998) and Iyer et al. (1999) mentioned that

genes involved in a common process tend to be co-expressed. Co-expression analyses have therefore been utilized for predicting functional relatedness such as the commonality of metabolic pathways (Li 2002), protein complexes (Geisler-Lee et al. 2007), and stress responses (Clifton et al. 2005).

The expression of soybean genes has been analyzed since the release of the genome draft. Haerizadeh et al. reported genome-wide analysis of gene expression in soybean shoot apical meristem (2009a) and pollen (2009b). Wong et al. (2009) adopted a clustering approach to predict co-expressed genes using a time-serial gene expression dataset in shoot apical meristem. Among such clustering approaches, a network approach has been used to provide systematic information on gene-to-gene associations based on microarray assays performed in various experiments (Manfield et al. 2006; Obayashi et al. 2009). Information on experiments contributing to such associations is useful for our understanding of functional relatedness between genes in biological processes.

We analyzed publicly available DNA microarray datasets of soybean and constructed a soybean gene co-expression database for genome-wide prediction of gene function on the basis of co-expression network analysis. We obtained 2995 publicly available DNA microarray

---

datasets (Affymetrix Soybean GeneChip®) from GEO. Using a co-expression network approach, we analyzed the datasets and extracted 'co-expression modules', which comprise genes that are tightly interconnected to each other on the basis of co-expression relationships. Co-expression modules were associated with descriptions in the experiments contributing to their intra-modular connections. Here, we illustrate two modules that are composed of genes involved in isoflavonoid biosynthesis and secondary cell wall biosynthesis, respectively. Our database is available at http://webs2.kazusa.or.jp/kagiana/cop/.

We downloaded gene expression datasets of 2995 Affymetrix Soybean GeneChips®, which are composed of 61 035 probes based on expressed sequence tags, from the GEO database. The data files were processed using Bioconductor 2.3.13 on R version 2.8.1 to obtain a set of text-formatted gene expression data. The data were standardized between microarray chips as well as between probes, referred to as 'Z score' in pages for co-expression modules, and cosine correlation coefficients between all pairs of probes were calculated.

As functional annotations of soybean genes comprising a co-expression module are required for functional characterization of the module, we adopted those of Arabidopsis genes which are homologous to soybean genes. As an index of homology, we defined a harmonic mean of mutual values of identities in base sequences, referred to as the Homology F-measure (HF). We provide functional annotations of Arabidopsis genes with the highest HF values to soybean genes in the module on the basis of analysis using BLAST.

Using publicly available DNA microarray datasets of soybean genes, we constructed a database for soybean co-expression analysis, which is implemented under the CoP database management system (http://webs2.kazusa.or.jp/kagiana/cop/) that allows users to perform co-expression analyses of Arabidopsis, poplar (Ogata et al. 2009a), and soybean and retrieve biological processes based on Gene Ontology (Gene Ontology Consortium 2008). By performing a query of a gene of interest, the database displays a page providing descriptions on the co-expression module comprising genes co-expressed with the query gene.

To evaluate functional relatedness of co-expressed genes, we adopted the network F-measure index (NF), which was introduced as the NC index by Ogata et al. (2009b).

In the portal page of the database (Figure 1), steps to input required items are as follows:

I) input a query word e.g., a probe identifier, gene identifier, Arabidopsis gene identifier (AGI code), or keyword included in functional description terms; II) select a plant organisms e.g., "Glycine max (soybean), Affymetrix 61k probes, 3k assays all"; III) select an

Figure 1.   Retrieval form of the CoP database. The retrieval steps are as follows: I) input a query word e.g., an AGI code and a microarray probe name; II) select a "Glycine max (soybean) Affymetrix 61k probes, 3k assays all"; III) select "Confeito (co-expression analysis)"; IV) skip this step; and V) click the "Submit" button. When a single candidate gene is found, a page including descriptions of a co-expression module that is composed of genes co-expressed with the gene is directly displayed. Otherwise, a page of a list of gene identifiers that are linked to pages including the descriptions of co-expression modules is displayed.

information type e.g., "Confeito (co-expression analysis)"; IV) skip this step to obtain co-expressed genes; and then V) click the submit button.

A page produced by the query is comprised of parts representing descriptions of a co-expression module (Figure 2). The first part includes the query gene identifier, tightness index of the module (referred to as the Network F-measure; NF), and module size (the number of genes). Next, the 'Descriptions' part is composed of probe identifiers, representative public identifiers, and information of the homologous Arabidopsis genes; i.e., AGI code, HF, gene names, short descriptions on their functions, and GO biological processes. In the third part, the descriptions of 'Specific Experiments', in which the co-expressed genes are specifically expressed, are composed of standardized scores of the genes ("Z score"), sample names,

Confeito : Gma.8449.1.S1_at

NF: 0.503, 18 kernel genes

Descriptions

Probe name	Repr. public ID	AGI code	HF	Gene name	Short description	Biological process
Gma.5491.1.S1_at	AW309940	At5g51105	0.020	-	-	biological_process_unknown
Gma.5491.2.S1_a_at	BE425401	At5g59570	0.029	-	-	regulation of transcription
Gma.8449.1.S1_at	AF002258.1	At1g65060	0.124	4CL3	4CL3 (4-coumarate:CoA ligase 3)	response to UV
Gma.8449.1.S1_s_at	AF002258.1	At1g65060	0.124	4CL3	4CL3 (4-coumarate:CoA ligase 3)	response to UV
Gma.9013.1.S1_at	AF243368.1	At1g74590	0.067	ATGSTU10	ATGSTU10 (Arabidopsis thaliana Glutathione S-transferase (class tau) 10)	toxin catabolic process
Gma.9013.1.S1_x_at	AF243368.1	At1g74590	0.067	ATGSTU10	ATGSTU10 (Arabidopsis thaliana Glutathione S-transferase (class tau) 10)	toxin catabolic process
GmaAffx.18616.1.S1_at	BE021154	At4g34540	0.061	-	isoflavone reductase family	regulation of nitrogen utilization
GmaAffx.43343.3.S1_s_at	AI496624	At2g27503	0.036	-	-	-
Gma.13342.1.A1_at	AW349994	At4g13660	0.037	ATPRR2,PRR2	pinoresinol-lariciresinol reductase, putative	lignan biosynthetic process
GmaAffx.89508.1.A1_s_at	CK606172	At3g10340	0.045	PAL4	phenylalanine ammonia-lyase, putative	L-phenylalanine catabolic process
GmaAffx.90356.1.S1_s_at	CF806081	At4g13660	0.053	ATPRR2,PRR2	pinoresinol-lariciresinol reductase, putative	lignan biosynthetic process
GmaAffx.91013.1.S1_s_at	CF806738	At5g51105	0.030	-	-	biological_process_unknown
Gma.15492.1.S1_s_at	AF276302.1	At1g52510	0.018	-	hydrolase, alpha/beta fold family	-
Gma.2308.1.S1_s_at	AF022462.1	At1g67195	0.018	-	-	-
Gma.2555.1.S1_s_at	BE819851	At2g38740	0.172	-	haloacid dehalogenase-like hydrolase family	metabolic process
Gma.4300.1.S1_s_at	S46989.1	At5g13930	0.316	ATCHS,CHS,TT4	ATCHS/CHS/TT4 (CHALCONE SYNTHASE)	chalcone biosynthetic process
Gma.4300.3.S1_s_at	AI443042	At5g13930	0.147	ATCHS,CHS,TT4	ATCHS/CHS/TT4 (CHALCONE SYNTHASE)	chalcone biosynthetic process
Gma.4300.3.S1_x_at	AI443042	At5g13930	0.147	ATCHS,CHS,TT4	ATCHS/CHS/TT4 (CHALCONE SYNTHASE)	chalcone biosynthetic process

Specific Experiments

The specific experiments, where these genes are specifically expressed on the basis of their z-scores, are listed in the below table. Each experiment is directly hyperlinked to the page in the GEO database.

Z score	Sample name	Experiment ID	Description	Experiment title
3.0	GSM293550:64_2021.2A,R275U, Experimental rep 2	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.9	GSM292799:42_2160.2D,R076P, Experimental rep 2	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.8	GSM171695:32_W.K.I.5, Experimental rep W	GSE7124	Description	Plant and pathogen gene expression during infection by P.sojae of 8 soybean cultivars varying in quantitative resistance
2.6	GSM171606:12_V.I.M.5, Experimental rep V	GSE7124	Description	Plant and pathogen gene expression during infection by P.sojae of 8 soybean cultivars varying in quantitative resistance
2.6	GSM171710:6_V.L.M.5, Experimental rep V	GSE7124	Description	Plant and pathogen gene expression during infection by P.sojae of 8 soybean cultivars varying in quantitative resistance
2.6	GSM171628:17_Y.A.I.3, Experimental rep Y	GSE7124	Description	Plant and pathogen gene expression during infection by P.sojae of 8 soybean cultivars varying in quantitative resistance
2.6	GSM292507:23_C140.1H,R9H3P, Experimental rep 1	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.6	GSM171649:21_Z.K.M.5, Experimental rep Z	GSE7124	Description	Plant and pathogen gene expression during infection by P.sojae of 8 soybean cultivars varying in quantitative resistance
2.5	GSM292739:83_C175.1J,R9J3L, Experimental rep 1	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.5	GSM292839:19_C199.1LV71U, Experimental rep 1	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.5	GSM293332:48_2019.2A,R192U, Experimental rep 2	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.5	GSM294166:2_2384.2H,R255P, Experimental rep 2	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.5	GSM171662:25_Y.K.M.5, Experimental rep V	GSE7124	Description	Plant and pathogen gene expression during infection by P.sojae of 8 soybean cultivars varying in quantitative resistance
2.5	GSM293737:70_2126.2C,R103P, Experimental rep 2	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.5	GSM293540:55_2016.2A,R250P, Experimental rep 2	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.4	GSM293223:12_1179.1P,R110L, Experimental rep 1	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.4	GSM171682:2_V.A.M.5, Experimental rep V	GSE7124	Description	Plant and pathogen gene expression during infection by P.sojae of 8 soybean cultivars varying in quantitative resistance
2.3	GSM294235:84_2383.2H,R255L, Experimental rep 2	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.3	GSM292521:36_1291.1H,R185L, Experimental rep 2	GSE11611	Description	Combined gene expression and QTL analysis of soybean quantitative resistance to Phytophthora sojae
2.3	GSM244911:40_SLO.M2W3.O8J, Experimental rep 3	GSE9687	Description	Expression patterns in time and space during P.sojae infection of soybean cultivars differing in quantitative resistance

Figure 2. Descriptions of a co-expression module. Functional descriptions of a co-expression module in which genes may be involved in isoflavonoid biosynthesis. The "Descriptions" includes the descriptions of the genes. The "HF" column represents homology indices of soybean genes with Arabidopsis genes in a co-expression module. In this example, the HF values are not high (<0.1) except for "4CL3" (0.124) and "ATCHS, CHS, TT4" (0.316). The "Short description" column, which shows a functional description of homologous Arabidopsis genes, includes "4CL3", "isoflavone reductase family", "pinoresinol-lariciresinol reductase, putative", "phenylalanine ammonia-lyase, putative", and "ATCHS/CHS/TT4 (CHALCONE SYNTHASE)". In the "Specific Experiments" part, the "Z score" column represents standardized expression values averaged in the genes: the higher value (e.g., >3) represents more specific expression.

experimental identifiers of GEO, links to detailed descriptions of experiments, and experiment titles. The descriptions of this part support users for understanding a biological process including the module.

Practical steps to retrieve a co-expression module are as follows: in the portal site (http://webs2.kazusa.or.jp/kagiana/cop/),

I)   input "At1g65060" as an example of a query term,
II) select "Glycine max (soybean), Affymetrix 61k probes, 3k assays all",
III) select "Confeito (co-expression analysis)",
IV) skip this selection,
V)   click the "Submit" button, and then in the page of a list of probe identifiers,
VI) click "Gma.8449.1.S1_at", whose Arabidopsis homologous gene is named "4CL3 (4-coumarate: CoA ligase 3)" to display a summarized description of the co-expression module.

Figure 2 represents the page displayed by the query. In the page, descriptions of a co-expression module comprising 18 genes that are co-expressed with the query gene are included. The "Descriptions" part of this module includes functional descriptions related to isoflavonoid biosynthesis i.e., "PAL4 (phenylalanine ammonia-lyase, putative)", "ATCHS/CHS/TT4 (CHALCONE SYNTHASE)", and "isoflavone reductase family". The HF indices of these genes are low (<0.1) except for *4CL3* and *CHS*, indicating that the biological process in which the genes are involved has an expression regulatory system different from the similar process in Arabidopsis.

When performing query of "At4g18780", encoding "ATCESA8 (CELLULOSE SYNTHASE 8)", the co-expression module for the query gene includes genes involved in secondary cell wall biosynthesis; i.e., "CESA4 (CELLULOSE SYNTHASE 4)", "COBL4/IRX6 (COBRA-LIKE4)", "GAUT12/IRX8/LGT6 (GALACTURONOSYLTRANSFERASE 12)", and "IRX3 (IRREGULAR XYLEM 3, MURUS 10)". Brown et al. (2009) revealed that "GUT1" has an essential role in glucuronoxylan biosynthesis and referred to it as *IRX10*-like, which is closely related to *IRX10* (86% amino acid identity). In contrast to the module co-expressed with *4CL3*, the module co-expressed with *ATCESA8* is mainly composed of genes with high HF scores (>0.1), indicating that members of the co-expression module are similar to those in Arabidopsis.

Our database provides information on soybean co-expression modules for predicting gene function and contributing toward understanding of biological processes. By performing a query with an identifier of a gene of interest or a keyword related to a gene of interest, users can retrieve a co-expression module that is composed of genes co-expressed with the query gene.

The database provides a list of co-expressed modules with tight intra-modular connections in descending order of their NF values. By using the term "confeito" as a query word, a list in which probe identifiers are clickable to jump to pages for a co-expression module including the identifier, is displayed. This approach allows users to access a page for a co-expression module with tight intra-modular connections without knowledge of gene functions or biological processes.

The present database provides information on experiments in which genes in a co-expression module are specifically expressed. In public databases for plant co-expression analyses such as ATTED-II (Obayashi et al. 2009; http://atted.jp/), CSB.DB (Steinhauser et al. 2004; http://csbdb.mpimp-golm.mpg.de/), and ACT (Manfield et al. 2006; http://www.arabidopsis.leeds.ac.uk/act/), co-expressed genes are extracted using gene-to-gene correlation data based on publicly available DNA microarray datasets. Such correlation data reflect strong influence by experiments in which genes are specifically expressed. Information on such experiments helps users characterize genes whose functions are unrevealed. In particular, co-expression analysis for soybean, whose genes are required for characterizing their functions through a systematic approach, may benefit from information from the experiments.

Our database provides information on the homology of soybean genes to Arabidopsis genes in a co-expression module for understanding the conservation of biological processes between plants. In the previous section, we illustrated two examples of co-expression modules with different homologies with respect to Arabidopsis genes i.e., the co-expression modules involved in isoflavonoid biosynthesis and secondary cell wall biosynthesis show low- and high-level homology to Arabidopsis genes, respectively. The difference in homology in the co-expression modules may result from the commonality, or lack thereof, in biological processes among plants.

By combining transcriptome data with that of genome, proteome, and interactome, gene co-expression may be better understood. The TIGR Plant Transcript Assemblies database and the GeneChip Oncology Database (Pertea et al. 2003; http://combio.dfci.harvard.edu/tgi/plant.html) provide genomic data of plants (Childs et al. 2007; http://plantta.jcvi.org/). Arabidopsis omics data including that of the genome, transcriptome, and proteome are available at the TAIR database (Swarbreck et al. 2008; http://arabidopsis.org/). The combination of omics data in soybean provides useful information for understanding the functionality of co-expression modules and regulation of such modules in legumes.

Acknowledgements

References

Alvord WG, Roayaei JA, Quiñones OA, Schneider KT (2007) A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R. *Brief Bioinformatics* 8: 415–443

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815

Brown DM, Zhang Z, Stephens E, Dupree P, Turner SR (2009) Characterization of IRX10 and IRX10-like reveals an essential role in glucuronoxylan biosynthesis in Arabidopsis. *Plant J* 57: 732–746

Cheng KCC, Strömvik MV (2008) SoyXpress: A database for exploring the soybean transcriptome. *BMC Genomics* 9: 368

Cheng KC, Beaulieu J, Iquira E, Belzile FJ, Fortin MG, Strömvik MV (2008) Effect of transgenes on global gene expression in soybean is within the natural range of variation of conventional cultivars. *J Agric Food Chem* 56: 3057–3067

Childs KL, Hamilton JP, Zhu W, Ly E, Cheung F, Wu H, Rabinowicz PD, Town CD, Buell C, Chan AP (2007) The TIGR Plant Transcript Assemblies database. *Nucl Acids Res* 35: D846–D851

Clifton R, Lister R, Parker KL, Sappl PG, Elhafez D, Millar AH, Day DA, Whelan J (2005) Stress-induced co-expression of alternative respiratory chain components in *Arabidopsis thaliana*. *Plant Mol Biol* 58: 193–212

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95: 14863–14868

Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M (2007) A predicted interactome for Arabidopsis. *Plant Physiol* 145: 317–329

Gene Ontology Consortium (2008) The Gene Ontology project in 2008. *Nucl Acids Res* 36: D440–D444

Haerizadeh F, Wong CE, Singh MB, Bhalla PL (2009a) Genome-wide analysis of gene expression in soybean shoot apical meristem. *Plant Mol Biol* 69: 711–727

Haerizadeh F, Wong CE, Bhalla PL, Gresshoff PM, Singh MB (2009b) Genomic expression profiling of mature soybean (*Glycine max*) pollen. *BMC Plant Biol* 9: 25

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436: 793–800

Iyer VR, Eisen MB, Ross DT, Schler G, Moore T, et al. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* 283: 83–87

Li K (2002) Genome-wide coexpression dynamics: Theory and application. *Proc Natl Acad Sci USA* 99: 16875–16880

Manfield IW, Jen CH, Pinney JW, Michalopoulos I, Bradford JR, Gilmartin PM, Westhead DR (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. *Nucl Acids Res* 34: W504–W509

Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucl Acids Res* 37: D987–D991

Ogata Y, Suzuki H, Shibata D (2009a) A database for poplar gene co-expression analysis for systematic understanding of biological processes, including stress responses. *J Wood Sci* 55: DOI: 10,1007/s10086-009~1058-9

Ogata Y, Sakurai N, Suzuki H, Aoki K, Saito K, Shibata D (2009b) The prediction of local modular structures in a co-expression network based on gene expression datasets. *Genome Informatics* (in press)

Panthee DR, Yuan JS, Wright DL, Marois JJ, Mailhot D, Stewart CNJ (2007) Gene expression analysis in soybean in response to the causal agent of Asian soybean rust (*Phakopsora pachyrhizi* Sydow) in an early growth stage. *Funct Integr Genomics* 7: 291–301

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, et al. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19: 651–652

Steinhauser D, Usadel B, Luedemann A, Thimm O, Kopka J (2004) CSB.DB: a comprehensive systems-biology database. *Bioinformatics* 20: 3647–3651

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, et al. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucl Acids Res* 36: D1009–D1014

Umezawa T, Sakurai T, Totoki Y, Toyoda A, Seki M, et al. (2008) Sequencing and analysis of approximately 40 000 soybean cDNA clones from a full-length-enriched cDNA library. *DNA Res* 15: 333–346

Wong CE, Singh MB, Bhalla PL (2009) Molecular processes underlying the floral transition in the soybean shoot apical meristem. *Plant J* 57: 832–845