CoP: a database for characterizing co-expressed gene modules with biological information in plants

Yoshiyuki Ogata, Hideyuki Suzuki, Nozomu Sakurai and Daisuke Shibata*

Department of Biotechnology Research, Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan

Received on November 12, 2009; revised on February 11, 2010; accepted on March 16, 2010

1 INTRODUCTION

Since the decoding of the Arabidopsis genome (Arabidopsis Genome Initiative, 2000), information on genome sequences and gene expression in plants has accumulated in public databases. Many microarray datasets assembled from various experiments (e.g. different tissues and chemical treatments), including the AtGenExpress datasets of Arabidopsis thaliana (thale cress) and seven crops, Glycine max (soybean), Hordeum vulgare (barley), Oryza sativa (rice), Populus trichocarpa (poplar), Triticum aestivum (wheat), Vitis vinifera (grape) and Zea mays (maize). There are several databases that are designed to extract co-expressed genes from plant microarray datasets and then provide biological information with individual genes in the co-expressed gene group, facilitating investigation of gene function (see the review of Usadel et al., 2009). Cross-species comparison of relevant co-expressed gene groups is also useful, as seen in the database GeneCAT (Mutwil et al., 2008) that provides comparative analysis of Arabidopsis and barley co-expressed genes on the basis of sequence similarity. Association of co-expressed gene modules intentionally with biological information such as gene ontology (GO) and metabolic pathways would be useful. However, to our knowledge, there is no such database available. Thus, we have developed the CoP database for associating co-expressed gene modules assembled from public large microarray datasets with biological information and comparing these modules across plant species to provide fundamental data for hypothesis generation of gene function. We apply the ‘condition-independent’ co-expression analysis to extract co-expressed modules from the whole set of microarray data, which gives only a single score, irrespective of tissue types and other experimental conditions, to the gene of interest. This approach is a suitable way to view general gene-to-gene relationships for the initial investigation of genes of interest (see the review of Usadel et al., 2009). In addition to the biological information such as GO terms and pathway names, the system provides direct links to public databases for further information on the co-expressed gene modules. The current database utilizes public microarray datasets from Arabidopsis thaliana (thale cress), Glycine max (soybean; Ogata et al., 2009a), Hordeum vulgare (barley), Oryza sativa (rice), Populus trichocarpa (poplar; Ogata et al., 2009b), Triticum aestivum (wheat), Vitis vinifera (grape) and Zea mays (maize).

2 DATA ACQUISITION

DNA microarray datasets were obtained from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) databases. We obtained CEL files or MA55-processed data files (Affymetrix GeneChip) for A.thaliana (5257 chips), G.max (2994), H.vulgare (403), O.sativa (432), P.trichocarpa (95), T.aestivum (368), V.vinifera (210) and Z.mays (263). The CEL files were processed using the Bioconductor package 2.3.13 with R version 2.8.1 to obtain MA55-processed text data. The text data for each plant were standardized and then used to calculate gene-to-gene correlation in gene expression profiles. As a measure of the correlation, we use cosine correlation coefficient because it is suitable to deal with datasets that appear to be more reliable in high positive expression values than in low ones in terms of the signal-to-noise ratio. See the concept text attached to the database for details (http://webs2.kazusa.or.jp/kajiana/cop0911/concept.html).

Biological information was obtained from datasets of GO biological processes for A.thaliana (March 24, 2009) from TAIR (http://arabidopsis.org/) and for O.sativa (March 6, 2009) from GRAMENE (http://www.gramene.org/). To assign plant genes to metabolic pathways, we obtained information on the assignment for Arabidopsis genes from KEGG PATHWAY (http://www.genome.jp/kegg/pathway.html) and KaPPA-View 4 (http://kpv.kazusa.or.jp/kpv4/; Tokimatsu et al., 2005).

*To whom correspondence should be addressed.
TAIR database (TAIR8_cds_20080412, April 12, 2008), the
The CoP portal site allows users to input query terms according
We performed homology search using NEC Homology
NF indices of all other modules.
webs2.kazusa.or.jp/kagiana/cop0911/pages/terms.html for details). Query terms include gene identifiers (e.g.
plant; (iii) selection of information type; and then (iv) selection
processes, and links to additional biological information and
public databases for each gene. By clicking on 'More genes', all
co-expressed genes in the module are presented in the table. The
third section presents a list of microarray experiments on genes in the
module showing specific expression. The fourth section represents
a comparison between co-expressed gene modules in two plants.
Funding: The New Energy and Industrial Technology Development
Organization (NEDO) program that is part of the 'Development of
Fundamental Technologies for Controlling the Material Production
Process of Plants' project.
Conflict of Interest: none declared.
REFERENCES
Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering
Mattick,J.M. et al. (2008) GeneCAT – novel webtools that combine BLAST and
Ohyama,T. et al. (2009) ATTED-II provides coexpressed gene networks for
Ogata,Y. et al. (2006a) A gene co-expression database for understanding biological
Ogata,Y. et al. (2006b) ATTED-II provides coexpressed gene networks for
Ogata,Y. et al. (2009a) A database for poplar gene co-expression analysis for systematic
Ogata,Y. et al. (2009b) PATHWAY is also represented with the harmonic mean index.

5 DATABASE DESIGN
5.1 The CoP portal site
The CoP portal site allows users to input query terms according to
the following steps: (i) input of query term; (ii) selection of
plant; (iii) selection of option types; and then (iv) selection of
options (see http://webs2.kazusa.or.jp/kagiana/cop911/pages/
terms.html for details). Query terms include gene identifiers (e.g.
an AGI code for Arabidopsis), probe names and all or part of a
gene name. When multiple hits are obtained for genes or probes for a
query, a list of these hits is displayed with corresponding links to
individual information pages. The complete list of the co-expressed
gene modules can be displayed in descending order of the tight
connectivity by entering the search string 'confeito' to the 'Gene,
co-expression'.

5.2 Information pages for co-expressed gene modules
The information page for each co-expressed gene module comprises
four sections. The first section provides information about the
connectivity and number of genes of the module. The Arabidopsis
page also includes a co-expressed gene network graph with the
query gene; i.e. genes are interconnected by their similarity in
expression profiles. In each network, nodes are categorized into four
functional types: transcription factor, binding protein, enzyme and
other proteins. The second section presents a table of functional
descriptions, including the associations of genes with biological
processes, and links to additional biological information and
User
Ogata,Y. et al. (2006a) A gene co-expression database for understanding biological
Ogata,Y. et al. (2006b) ATTED-II provides coexpressed gene networks for
Ogata,Y. et al. (2009a) A database for poplar gene co-expression analysis for systematic
Ogata,Y. et al. (2009b) PATHWAY is also represented with the harmonic mean index.

5 DATABASE DESIGN
5.1 The CoP portal site
The CoP portal site allows users to input query terms according to
the following steps: (i) input of query term; (ii) selection of
plant; (iii) selection of option types; and then (iv) selection of
options (see http://webs2.kazusa.or.jp/kagiana/cop911/pages/
terms.html for details). Query terms include gene identifiers (e.g.
an AGI code for Arabidopsis), probe names and all or part of a
gene name. When multiple hits are obtained for genes or probes for a
query, a list of these hits is displayed with corresponding links to
individual information pages. The complete list of the co-expressed
gene modules can be displayed in descending order of the tight
connectivity by entering the search string 'confeito' to the 'Gene,
co-expression'.

5.2 Information pages for co-expressed gene modules
The information page for each co-expressed gene module comprises
four sections. The first section provides information about the
connectivity and number of genes of the module. The Arabidopsis
page also includes a co-expressed gene network graph with the
query gene; i.e. genes are interconnected by their similarity in
expression profiles. In each network, nodes are categorized into four
functional types: transcription factor, binding protein, enzyme and
other proteins. The second section presents a table of functional
descriptions, including the associations of genes with biological
processes, and links to additional biological information and

User
Ogata,Y. et al. (2006a) A gene co-expression database for understanding biological
Ogata,Y. et al. (2006b) ATTED-II provides coexpressed gene networks for
Ogata,Y. et al. (2009a) A database for poplar gene co-expression analysis for systematic
Ogata,Y. et al. (2009b) PATHWAY is also represented with the harmonic mean index.

5 DATABASE DESIGN
5.1 The CoP portal site
The CoP portal site allows users to input query terms according to
the following steps: (i) input of query term; (ii) selection of
plant; (iii) selection of option types; and then (iv) selection of
options (see http://webs2.kazusa.or.jp/kagiana/cop911/pages/
terms.html for details). Query terms include gene identifiers (e.g.
an AGI code for Arabidopsis), probe names and all or part of a
gene name. When multiple hits are obtained for genes or probes for a
query, a list of these hits is displayed with corresponding links to
individual information pages. The complete list of the co-expressed
gene modules can be displayed in descending order of the tight
connectivity by entering the search string 'confeito' to the 'Gene,
co-expression'.