BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 9 2010, pages 1267–1268
doi:10.1093/bioinformatics/btq121

Gene expression

Advance Access publication March 19, 2010

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

Associate Editor: Trey Ideker

ABSTRACT

Summary: Using a large dataset (10 022 assays) obtained from public plant microarray databases, we developed the CoP database for associating co-expressed gene modules with biological information such as gene ontology terms and, if available, metabolic pathway names. The Confeito algorithm developed previously in our laboratory, which is suitable to calculate the interconnectivity between genes in co-expressed gene network, was applied to extract co-expressed gene modules. The database includes the gene modules for *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).

Availability: The CoP database is available at: http://webs2.kazusa.or.jp/kagiana/cop0911/

Contact: shibata@kazusa.or.jp

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* (Schmid *et al*., 2005), have been utilized to predict co-expressed genes for assigning them to metabolic pathways (Obayashi *et al*., 2009; Wei *et al*., 2006), transcriptional regulation (Hirai *et al*., 2007; Obayashi *et al*., 2009) and/or biological processes (Ogata and Shibata, 2009).

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 MAS5-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 MAS5-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/kagiana/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.

The dataset of *Arabidopsis* cDNA was obtained from the TAIR database (TAIR8_cds_20080412, April 12, 2008), the dataset of rice cDNA from RAP-DB (http://rapdb.dna.affrc.go.jp/; prediction_nuc.fa and rep_nuc.fa, May 16, 2007) cDNA and other plant EST datasets from the DFCI database (http://compbio.dfci.harvard.edu/tgi/; these latest version data were obtained in February 2009) for homology analysis.

3 DETECTION OF CO-EXPRESSED GENE MODULES

We extracted co-expressed gene modules with the Confeito algorithm that is designed to detect highly interconnected modules from co-expressed gene networks (Ogata *et al*., 2009c). The algorithm uses a novel network index NF for each module (see http://webs2.kazusa.or.jp/kagiana/cop0911/pages/terms.html for details). To help assess the modules statistically, we also provide the percentile score of the NF index of the module of interest to the NF indices of all other modules.

4 ASSOCIATION WITH BIOLOGICAL INFORMATION

We performed homology search using NEC Homology Searcher 4.3a software with the BLASTn algorithm (NEC Corporation, Tokyo, Japan). Each gene sequence of a plant was BLAST searched not only to all sequences of the plant to find paralogous genes within the species, but also to all sequences of other plant to find the best hit orthologous gene. These homologous gene sets deposited in the database are used to represent the genes homologous to the user's query gene. Sequence similarity is represented with the commonly used indexes, *E*-value and Bit score, and also with the harmonic mean of identity. To associate each co-expressed gene module with biological information, the system compares the gene members of the module to those assigned to a GO term of biological processes that has been defined by the GO consortium with a harmonic mean index (see the detail in the concept text). Information of GO term association is provided only for *Arabidopsis* and rice, as for these plants the GO terms are available with the gene codes used in the CoP database. Association with metabolic pathway names of the omics integration tool KaPPA-View 4 and biological pathways of the database KEGG 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 information type; and then (iv) selection of options (see http://webs2.kazusa.or.jp/kagiana/cop0911/pages/terms.html for details). Query terms include gene identifiers (e.g. an AGI code for *A.thaliana*), 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 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 plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Hirai,M.Y. *et al*. (2007) Omics-based identification of *Arabidopsis* Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proc. Natl Acad. Sci. USA*, **104**, 6478–6483.

Mutwil,M. *et al*. (2008) GeneCAT – novel webtools that combine BLAST and co-expression analyses. *Nucleic Acids Res.*, **36**, W320–W326.

Obayashi,T. *et al*. (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Res.*, **37**, D987–D991.

Ogata,Y. *et al*. (2009a) A gene co-expression database for understanding biological processes in soybean. *Plant Biotechnol.*, **26**, 503–507.

Ogata,Y. *et al*. (2009b) A database for poplar gene co-expression analysis for systematic understanding of biological processes, including stress responses. *J. Wood Sci.*, **55**, 395–400.

Ogata,Y. *et al*. (2009c) The prediction of local modular structures in a co-expression network based on gene expression datasets. *Genome Inform.*, **23**, 117–127.

Ogata,Y. and Shibata,D. (2009) Practical network approaches and biologic interpretations of co-expression analyses in plants. *Plant Biotechnol.*, **26**, 3–7.

Schmid,M. *et al*. (2005) A gene expression map of Arabidopsis thaliana development. *Nat. Genet.*, **37**, 501–506.

Tokimatsu,T. *et al*. (2005) KaPPA-View: a web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. *Plant Physiol.*, **138**, 1289–1300.

Usadel,B. *et al*. (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ.*, **32**, 1633–1651.

Wei,H. *et al*. (2006) Transcriptional coordination of the metabolic network in Arabidopsis. *Plant Physiol.*, **142**, 762–774.