Bioinformatics Advance Access published October 17, 2006

Extending MapMan: Application to Legume Genome Arrays

Nicolas Goffard and Georg Weiller*

ARC Centre of Excellence for Integrative Legume Research and Bioinformatics Laboratory, Genomic Interactions Group, Research School of Biological Sciences, Australian National University, GPO Box 475, Canberra, ACT 2601 Australia

Associate Editor: David Rocke

ABSTRACT

Motivation: Based on a gene classification into hierarchical categories ('BINs'), MapMan was originally developed to display *Arabidopsis thaliana* gene expression in a functional context. We have created a bioinformatics system to extend MapMan to any organism by using a new BIN structure based on the KEGG database. Gene sequences are assigned to this ontology by homology relationships in four reference databases: KEGG, COG, Swiss-Prot and Gene Ontology. We applied this system to tailor MapMan to the GeneChips of two model legumes, *Glycine max* and *Medicago truncatula*. We also developed a module to identify the most relevant pathways involved.

Availability: All mapping files, pathway pictures and the analysis method are available at: http://bioinfoserver.rsbs.anu.edu.au/

Contact: georg.weiller@anu.edu.au

1 INTRODUCTION

Microarrays enable us to study the expression of thousands of genes simultaneously, providing a comprehensive overview of the gene activities in a given tissue. Bioinformatics tools such as MapMan can display this data in a functional context (Thimm, et al., 2004; Usadel, et al., 2005). MapMan requires three types of information: 1) a hierarchical classification of genes (i.e. BINs), 2) images representing a functional context of these genes (e.g. metabolic pathways) and 3) experimental expression data. The transcriptional activities of the binned genes are then displayed on the images using various statistical representations. Although initially developed for *Arabidopsis thaliana* arrays, MapMan can be extended to other systems by assigning new sequences to their orthologs in the current *A. thaliana* BINs (Urbanczyk-Wochniak, et al., 2006). However, this approach is limited as sequence similarity may be marginal with organisms of interest.

We propose a new strategy to tailor MapMan to any organism by defining a generic BIN structure modelled on the Kyoto Encyclopedia of Genes and Genomes (KEGG) ontology (Kanehisa, et al., 2004). We converted KEGG pathway pictures for visualization in MapMan and created additional overview images. We also devised a method based on the hypergeometric distribution as developed in the BlastSets system (Barriot, et al., 2004) in order to identify the most relevant pathways affected in transcriptomic experiments.

This approach was applied to the Affymetrix GeneChips of two major legumes, *Glycine max* (soybean) and *Medicago truncatula* (barrel medic).

2 EXTENDING MAPMAN

2.1 Definition of a new BIN structure

We first defined a new BIN structure by converting the entire KEGG Orthology database (Release 36) into MapMan BINs, with BIN names corresponding to pathway description or KEGG Ortholog (KO) functions and BIN descriptions containing KO identifiers as well as the cross-links to Clusters of Orthologous Groups (COG), Enzyme Commission (EC) numbers and Gene Ontology (GO) provided by KEGG (Figure 1c). In the original KO database several KOs can share COG, EC or GO terms, indicating different subclasses of an enzyme or different subunits of a complex. In these cases BINs were further split according to COG, EC or GO information associated with the KO, increasing the depth of the hierarchy. In addition, a special 'catch-all' BIN 'Unclassified' was added.

2.2 Gene assignments

We created four databases for the Blast searches shown in Figure 1a. 345,817 sequences and their associated KO identifiers were extracted from the KO database (Kanehisa, et al., 2004) (Release 36). 77,084 sequences and their corresponding COG identifiers were downloaded from the COG database (Tatusov, et al., 2001). 201,594 sequences were retrieved from Swiss-Prot (Bairoch, et al., 2005) (Release 6.6), with 77,675 associated with an EC number in their description lines. Gene Ontology (Ashburner, et al., 2000) (release 2005-12) provided 55,454 sequences associated with GO identifiers. We used Blastx (Altschul, et al., 1990) to find the best matches (E-value < 10^{-8}) for each probe set consensus sequence (i.e. sequence derived from the most 5' to the most 3' position in the public Unigene cluster) of the Affymetrix GeneChips analysed. From these we extracted KO, COG, EC and GO identifiers to assign the probe set to the corresponding BIN (Figure 1b). If a Blast match provided no identifier, the sequence was assigned to the special BIN 'Unclassified with homolog'. All remaining probe sets

*To whom correspondence should be addressed.

were assigned to the BIN 'Unclassified without homolog'. Where a gene was assigned to several levels within a BIN sub-tree, only the deepest level was retained.

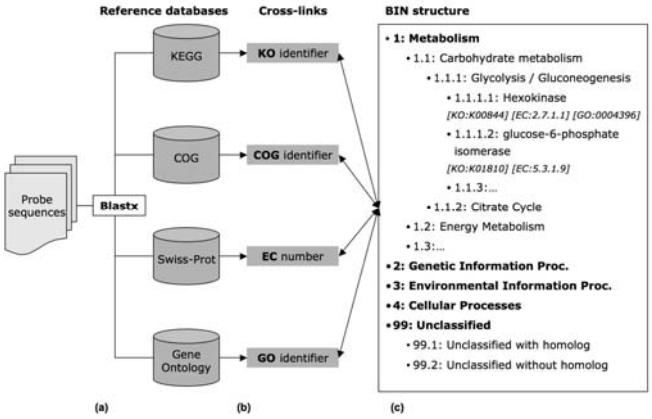

Fig. 1. The classification of new sequences is composed of 3 steps. First, using Blast, best matches are identified in reference databases for each sequence (a). Then, cross-links are extracted from the description of these database entries (b). Finally, sequences are assigned into the BIN Structure based on the relationship between cross-links and BINs (c).

The resulting mapping file can be imported into MapMan. We applied this procedure to the *Medicago truncatula* and *Glycine max* GeneChips (Irizarry, et al., 2003). As of September 2006, for *M. truncatula*, 13,322 of the 50,900 probe sets have been assigned to classified BINs, 15,990 to the 'Unclassified with homolog' BIN and 21,588 to the 'Unclassified without homolog' BIN. For the 37,618 *G. max* probe sets, the numbers are 9,842, 13,286 and 14,086 respectively.

2.3 Conversion of pathway pictures

For each biological pathway, KEGG provides a corresponding XML file describing the location within the image where each ortholog acts. We converted these to MapMan XML files indicating the location of the corresponding BINs instead. KEGG images can be converted to png format using ImageMagick (http://www.imagemagick.org). In summary, 151 XML files linked to pathway pictures have been converted, and new ones representing the top level of the ontology have been created for use in MapMan. Other pictures can be created and directly annotated with MapMan according to this new BIN structure.

3 IDENTIFYING SIGNIFICANT BINS

We developed a module to identify consistently over- or under-represented BINs in a submitted list of genes (e.g. differentially expressed genes). The query list of genes is compared to the composition of each BIN, including sub-BINs, in the mapping file. For each test, a *P*-value, representing the probability that intersection of the given list with the list of genes belonging to the given BIN occurs by chance, is calculated using the hypergeometric distribution. Because multiple hypothesis tests are performed, the *P*-value significant threshold is adjusted using a Bonferroni correction with a default cut-off of 0.05. Finally, this module returns the list of BINs over-represented in the set of submitted genes.

4 SUMMARY

We have created a system that can extend MapMan to any organism or GeneChip without being restricted to sequence similarity with *Arabidopsis*. In addition, we developed a web-based resource to identify significant BINs using a hypergeometric distribution model. While our BIN structure exceeds the original MapMan BIN structure for metabolic pathways, some other processes are less represented. We intend to further develop the system by extending both the BIN structure and the reference databases for Blast searches.

All resources are freely available at:
http://bioinfoserver.rsbs.anu.edu.au/

ACKNOWLEDGEMENTS

This study was funded by an Australian Research Council Centre of Excellence grant.

REFERENCES

Altschul, S. *et al.* (1990) Basic local alignment search tool., *J Mol Biol*, **215**, 403-410.

Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

Bairoch, A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*, **33**, D154-159.

Barriot, R. *et al.* (2004) New strategy for the representation and the integration of biomolecular knowledge at a cellular scale, *Nucleic Acids Res*, **32**, 3581-3589.

Irizarry, R.A. *et al.* (2003) Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Res*, **31**, e15.

Kanehisa, M. *et al.* (2004) The KEGG resource for deciphering the genome., *Nucleic Acids Res*, **32**, D277-280.

Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res*, **29**, 22-28.

Thimm, O. *et al.* (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes., *Plant J*, **37**, 914-939.

Urbanczyk-Wochniak, E. *et al.* (2006) Conversion of MapMan to Allow the Analysis of Transcript Data from Solanaceous Species: Effects of Genetic and Environmental Alterations in Energy Metabolism in the Leaf, *Plant Mol Biol*, **60**, 773-792.

Usadel, B. *et al.* (2005) Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses., *Plant Physiol*, **138**, 1195-1204.