BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 16 2003, pages 2158–2160
DOI: 10.1093/bioinformatics/btg291

GARBAN: genomic analysis and rapid biological annotation of cDNA microarray and proteomic data

Luis A. Martínez-Cruz[1,§], Angel Rubio[2,§], María L. Martínez-Chantar[1], Alberto Labarga[2], Isabel Barrio[2], Adam Podhorski[2], Victor Segura[2], José L. Sevilla Campo[2], Matías A. Avila[1] and Jose M. Mato[1,*]

[1]Department of Internal Medicine, Unit of Proteomics, Genomics and Bioinformatics, University of Navarra, Pamplona-31008, Spain and [2]Centro de Estudios e Investigaciones Tecnológicas de Guipúzcoa (CEIT), San Sebastián, Spain

Received on February 14, 2003; revised on May 9, 2003; accepted on May 16, 2003

ABSTRACT

Summary: Genomic Analysis and Rapid Biological ANnotation (GARBAN) is a new tool that provides an integrated framework to analyze simultaneously and compare multiple data sets derived from microarray or proteomic experiments. It carries out automated classifications of genes or proteins according to the criteria of the Gene Ontology Consortium at a level of depth defined by the user. Additionally, it performs clustering analysis of all sets based on functional categories or on differential expression levels. GARBAN also provides graphical representations of the biological pathways in which all the genes/proteins participate.

Availability: http://garban.tecnun.es

Contact: jmmato@unav.es

Large-scale gene expression studies are one of the most recent breakthroughs in experimental molecular biology (DeRisi *et al*., 1997; Eisen *et al*., 1998; Marton *et al*., 1998; Debouck and Goodfellow, 1999; Khan *et al*., 1999). These studies result in an unprecedented volume of data that represent a major challenge for the biologists. The organization of genes into relevant clusters depending on their differential expression pattern and the identification of what genes are expressed in a coordinated manner across a set of conditions are the most common methods used to facilitate microarray interpretation (for review see Tamames *et al*., 2002). A number of separate software systems individually address some of these needs, such as databases and applications for clustering and visualization of microarray data (Eisen *et al*., 1998; Spellman *et al*., 1998; Lukashin and Fuchs, 2001); Expression Profiler: http://ep.ebi.ac.uk/; (Tamayo *et al*., 1999; Lemkin *et al*., 2001; Sherlock and Fuchs, 2000; Dougherty *et al*., 2002; Diehn *et al*., 2003), or public databases that contain high-quality information for the cDNA contained therein (Diehn *et al*., 2003, http://info.med.yale.edu/microarray/; http://www.ncbi.nlm.nih.gov; http://www.ensembl.org; http://www.genetics.ucla.edu/microarray/Software.htm; http://www.ebi.ac.uk/microarray/; http://www.ncbi.nlm.nih.gov/Unigene; http://www.ncbi.nlm.nih.gov/dbEST/; http://www.geneontology.org/external2go/tigr2go). However, little attention has been paid to alternative methods performing automated classification of genes/proteins in terms of their biological function. Some browsers allow searches for ontological terms and identify gene associations for several organisms (http://www.geneontology.org/external2go/ec2go; http://www.ebi.ac.uk/ego/QuickGO; http://www.godatabase.org/dev; http://www.informatics.jax.org/searches/GO_form.shtml; Expression Profiler: http://ep.ebi.ac.uk/ Doniger *et al*., 2003), but are applicable only to individual genes/terms or in the best case to single data sets (Doniger *et al*., 2003). However, to our knowledge, when dealing with multiple data sets, there are no unified systems capable of combining all the information surrounding microarray or proteomic experimentation. Genome Analysis and Rapid Biological ANnotation (GARBAN) provides bioinformatic tools to account for this need and accelerates this process classifying each gene/protein according to the three functional categories of the Gene Ontology Consortium (GO; http://www.godatabase.org/dev/database). These categories are organized according to a hierarchical tree of parent/child relationships at a GO level defined by the user. A direct access to the corresponding information in NCBI

*To whom correspondence should be addressed.

§ The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

(http://www.ncbi.nlm.nih.gov), SWISS-PROT (http://www.ebi.ac.uk/swissprot/) or Ensembl (http://www.ensembl.org) is available for each entry.

SYSTEM OVERVIEW

The main distinction of our system with respect to other sources of microarray/proteomic analysis is the flexibility to analyse and compare multiple data sets simultaneously with bioinformatic tools that link expression data to GO classifications. Additionally, it is implemented with software to carry out clustering analysis on the basis of GO categories. This may help to start making connections between biological processes previously thought to be unrelated. The features implemented in GARBAN are: (a) classification and comparison of multiple data sets according to GO criteria; (b) SOM clustering based on functional criteria; (c) SOM clustering based on differential expression patterns; (d) automated BLAST search for expressed sequence tags (ESTs) and genes or proteins of unknown function against daily updated databases (many cDNA microarray chips contain ∼50% ESTs without known function (http://www.affymetrics.com); (e) rapid identification of user-defined genes/proteins across multiple data sets (i.e. useful to search for known markers among multiple samples in pharmacokinetical studies); (f) graphical representation of metabolic pathways affected in the experiments.

GARBAN has been programmed in the PHP language (http://www.php.net) with data being stored in a relational database (http://www.mysql.com) and communicated to the user through the Apache Webserver (http://httpd.apache.org). Where needed, the user interface employs Java and JavaScript in addition to plain HTML. The system is built on a Dell 2650 PC server computer, Dual Xeon 2.4 GHz, with 2 GB RAM memory and 146 GB hard disk SCSI-RAID 1. Complete genome and proteome sets for human and mouse organisms have been assembled from the SPTR (SWISS-PROT + TrEMBL + TrEMBL-new) database (http://www.ebi.ac.uk/swissprot/), and from Ensembl (http://www.ensembl.org), NCBI Genebank (http://www.ncbi.nlm.nih.gov/), dbEST (http://www.ncbi.nlm.nih.gov/dbEST/), InterPro (Apweiler *et al*., 2001) and TIGR (http://www.geneontology.org/external2go/tigr2go) databases. The LocusLink (Pruitt and Maglott, 2001), GO (http://www.godatabase.org/dev/database) and GO–Mouse Genome Databases (http://www.informatics.jax.org/mgihome/) assignments have been utilized to assign GO terms to proteins in SWISS-PROT and TrEMBL (http://www.ebi.ac.uk/swissprot/) and to InterPro domains and families (Apweiler *et al*., 2001). Databases are updated weekly and matches recalculated. In doing this, all the data are synchronized, ensuring that all information in the analysis database points to the most recent versions of the underlying databases. SOM clustering is carried out with the SOM PAK software package (Kohonen, 2001).

ACKNOWLEDGEMENTS

We thank J.M.Bastero for his continuous support. This work was supported by grants SAF 99/0038 and SAF 2002-00168 from MCyT and ROI AA-12677 from the NIAAA.

REFERENCES

Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al*. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.

Debouck,C. and Goodfellow,P.N. (1999) DNA microarrays in drug discovery and development. *Nat. Genet.*, **21**(Suppl. 1), 48–50.

Diehn,M., Sherlock,G., Binkley,G., Jin,H., Matese,J.C., Hernandez-Boussard,T., Rees,C.A., Cherry,J.M., Botstein,D., Brown,P.O. and Alizadeh,A.A. (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31**, 219–223

Dougherty,E.R., Barrera,J., Brun,M., Kim,S., Cesar,R.M., Chen,Y., Bittner,M. and Trent,J.M. (2002) Inference from clustering with application to gene-expression microarrays. *J. Comput. Biol.*, **9**, 105–126.

Doniger,S.W., Salomonis,N., Dahlquist,K.D, Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.*, **4**, R7.

DeRisi,J.L., Iyer,V. and Brown,P. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Khan,J., Saal,L.H., Bittner,M.L., Chen,Y., Trent,J.M. and Meltzer,P.S. (1999) Expression profiling in cancer using cDNA microarrays. *Electrophoresis*, **20**, 223–229.

Kohonen,T. (2001) *Self-Organizing Maps. Springer Series in Information Sciences*, Vol. 30, Springer-Verlag, Berlin-Heidelberg.

Lemkin,P.F., Thornwall,G.C., Walton,K.D. and Hennighausen,L. (2000) The microarray explorer tool for data mining of cDNA microarrays: application for the mammary gland. *Nucleic Acids Res.*, **28**, 4452–4459.

Lukashin,A.V. and Fuchs,R. (2001) Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, **17**, 405–414.

Marton,M.J., DeRisi,J.L., Bennett,H.A., Iyer,V.R., Meyer,M.R., Roberts,C.J., Stoughton,R., Burchard,J., Slade,D., Dai,H. *et al*. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.*, **4**, 1293–1301.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

Sherlock,G. (2000) Analysis of large-scale gene expression data. *Curr. Opin. Immunol.*, **12**, 201–205.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.*, **9**, 3273–3297.

Tamames,J., Clark,D., Herrero,J., Dopazo,J., Blaschke,C., Fernandez,J.M., Oliveros,J.C. and Valencia,A. (2002) Bioinformatics methods for the analysis of expression arrays: data clustering and information extraction. *J. Biotechnol.*, **98**, 269–283.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.