BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 13 2005, pages 2957–2959
doi:10.1093/bioinformatics/bti467

Sequence analysis

DIG—a system for gene annotation and functional discovery

Mark DeLong[1,2,*], Guang Yao[1,2], Quanli Wang[1,3], Adrian Dobra[1,2,3], Esther P. Black[1,2], Jeffrey T. Chang[1,2], Andrea Bild[1,2], Mike West[1,3], Joseph R. Nevins[1,2] and Holly Dressman[1,2]

[1]Duke Institute for Genome Sciences and Policy and [2]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC 27710, USA and [3]Institute of Statistics and Decision Sciences, Duke University, Durham, NC, USA, 27708

Received on August 25, 2004; revised and accepted on April 23, 2005
Advance Access publication May 3, 2005

ABSTRACT

Summary: We describe a database and information discovery system named DIG (Duke Integrated Genomics) designed to facilitate the process of gene annotation and the discovery of functional context. The DIG system collects and organizes gene annotation and functional information, and includes tools that support an understanding of genes in a functional context by providing a framework for integrating and visualizing gene expression, protein interaction and literature-based interaction networks.

Availability: The DIG system is freely accessible at http://dig.cagp.duke.edu/

Contact: mdelong@cgt.duhs.duke.edu

Supplementary information: http://data.cgt.duke.edu.dig

Genome-scale molecular profiling studies have been applied to an increasing number of applications addressing a wide variety of problems in biology, including subtle variation in tumors that define sub-types of significant clinical importance (Alizadeh *et al*., 2000; Bhattacharjee *et al*., 2001; Golub *et al*., 1999; Perou *et al*., 2000; Pomeroy *et al*., 2002; Ramaswamy *et al*., 2001; Shipp *et al*., 2002; Singh *et al*., 2002). Related studies with analysis methods aiming to predict phenotypes of unknown samples based on training data illustrate the great potential for genome-scale gene expression information in prognostic applications (Huang *et al*., 2003; Pittman *et al*., 2004; van de Vijver *et al*., 2002; van'T Veer *et al*., 2002; West *et al*., 2001). One key need is for integrated data exploration tools to access and explore biological information on genes playing roles in these analyses. To this end, we have developed an information resource—DIG (Duke Integrated Genomics)—that aims to assist in the discovery of functional context for genes identified in expression (and other) experiments. DIG provides tools that (1) integrate information about genes from different databases, (2) enrich Medline searches for relevant citations through custom queries, (3) organize datasets in shared workspaces and (4) link to graphical displays for visualization. The system is freely publicly accessible (https://dig.cagp.duke.edu/).

DIG links a user-defined list of genes to diverse databases to provide a comprehensive summary of known information about any given gene or pathway. The user can upload lists of genes to DIG through a web interface (Fig. 1A). DIG can accept GenBank accession numbers or gene or protein symbols. DIG then produces summaries of the sources of information via cross-references to external databases including genetic information (LocusLink, Homologene), gene function (Gene Ontology), involvement in pathways (BioCarta, KEGG, GenMAPP), relevance to disease (OMIM) and literature (Medline). Figure 1B gives an example for genes identified in a breast cancer ER (estrogen receptor) analysis.

DIG also provides tools for searching Medline to identify the co-occurrence of gene citations in published literature; this often reflects a biological relationship (Jenssen *et al*., 2001). Given a collection of genes identified in a profiling experiment, DIG first searches for all pairwise combinations of genes to identify relationships that may facilitate an understanding of the list (Fig. 1C). This can reveal processes that may not be obvious from independent queries for genes. DIG also performs pairwise searches of genes against user-defined terms to identify citations relevant to a specific aspect of gene function, such as involvement in a particular disease. The output is a list of titles along with links to the PubMed abstract. These searches make use of aliases of the gene names found in LocusLink. The search list, including the aliases, is generated automatically when the query is performed.

DIG provides additional access to three sources of networks to extend the analysis of a given gene through the generation of networks: large-scale gene expression datasets, protein interaction datasets and the published biological literature. Although the source of information in each case is distinct, the logic for identifying the relevant information is similar, as is, the mechanism for presenting the information. First, gene expression datasets provide the opportunity to identify functionally related genes based on statistical models of patterns of association in expression. We have developed such an approach using large-scale, sparse graphical models (Dobra *et al*., 2003; Jones *et al*., 2004, http://ftp.isds.duke.edu/WorkingPapers/04-06.html; Rich *et al*., 2005). This is a powerful statistical approach that has the potential to substantially advance our ability to identify functional relationships since it does not depend only on similarity of expression patterns; rather, this is a full statistical model that assesses the partial associations between one gene and another in the context of all others via sparse regression analysis, as examples in the references listed above illustrate.

Second, protein–protein interactions can represent enzyme–substrate interactions, regulatory interactions or the interaction of structural components within the cell. The ability to generate

*To whom correspondence should be addressed.

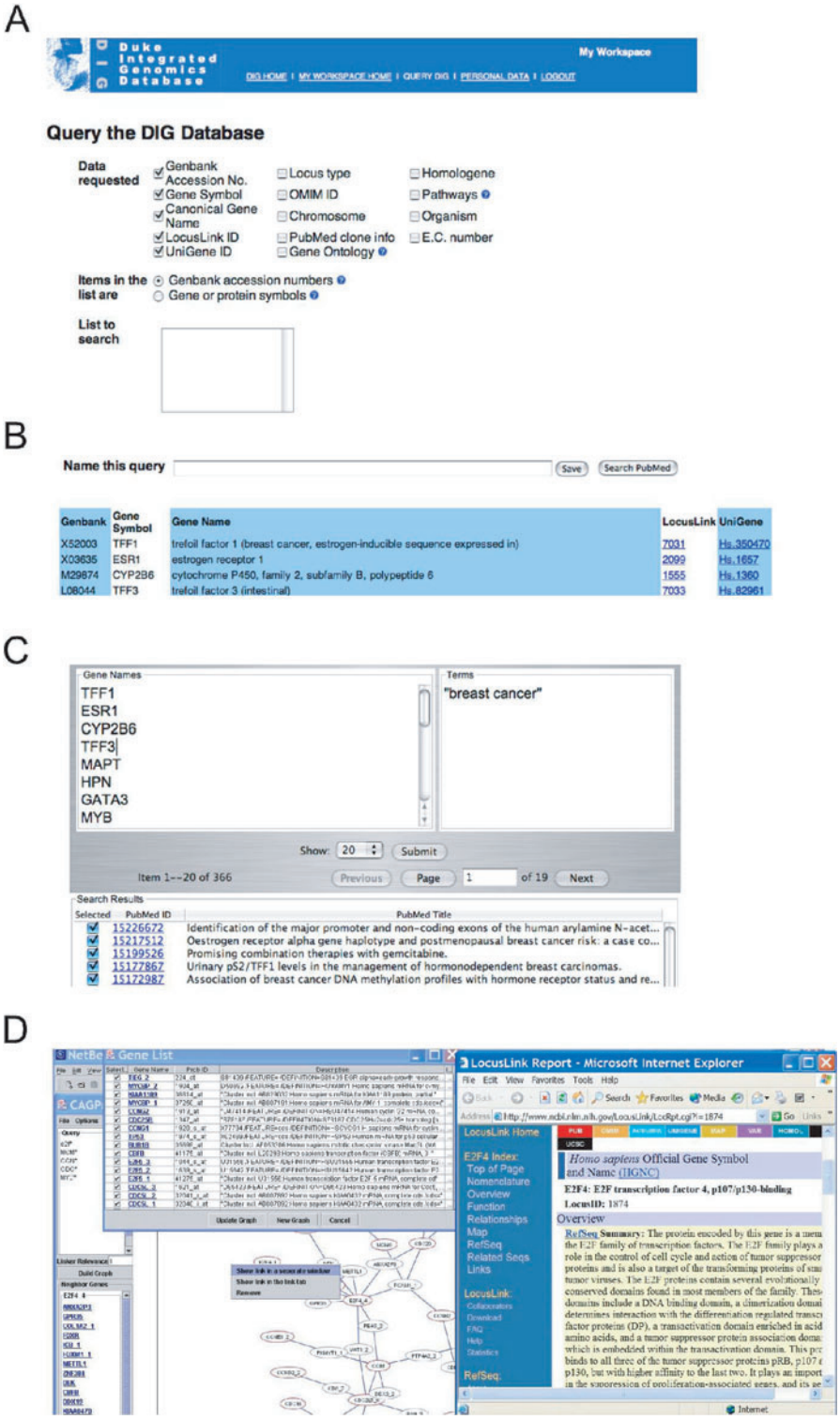

Fig. 1. The DIG system. (**A**) The DIG home page. The screenshot depicts the data entry page with selections for initial outputs. (**B**) Gene annotation in the DIG system. The screenshot depicts a gene annotation output, providing links to other data sources (pathways, LocusLink, UniGene). (**C**) The screenshot depicts the search page created when the 'Search PubMed' button on the DIG main page is clicked. The gene list is loaded into the gene search box. Clicking 'submit' launches a pairwise search of PubMed in which all genes in the list are searched against each other. The output shown at the bottom is the list of titles from the search with the associated link to the abstract. Each gene in the list is searched against any term(s) entered in the right-hand box (the example here is 'breast cancer'). (**D**) The screenshot depicts the GraphExplore interface, including the query genes, a graphical output of expression relationships and the hotlinks to LocusLink.

networks of these interactions ($A + B$ and $B + C$ equals $A + C$), rather than merely identifying the direct partners, has the potential to extend the understanding of functional networks beyond the immediate interacting proteins. The development of efficient, high-throughput methods for identifying protein interactions based on the yeast two-hybrid assay has facilitated the development of databases that describe these interactions.

Third, the published literature provides the opportunity to relate one gene to another—identifying functional relationships that will form the basis for establishing networks. The ability to identify publications that describe gene interactions or functional relationships between two or more gene products provides an opportunity to improve the understanding of expression profiling experimental results. DIG draws on a local copy of Medline to speed the processes of access, indexing, interrogation and searching. Since much of the basis of DIG is currently gene-centric, DIG has sub-indices containing instances of gene and protein names, and their variants over time, in the published literature. This information can represent gene and protein relationships within Medline in a manner similar to the representation of gene expression networks.

To manage and access the complexity inherent in the data that is rendered into networks, DIG has a core visualization facility (GraphExplore) that allows exploration of a visualized network with links to related online resources (Fig. 1D). GraphExplore is embedded as a core component for visualization within DIG, as well as acting as a stand-alone graphical visualization and exploration tool (Q. Wang, G. Yao, J.R. Nevins and M. West, submitted for publication, http://graphexplore.cgt.duke.edu). Users may create sub-network graphs by querying for lists of genes defined by statistically or biologically defined models. In addition to visualizing the network, GraphExplore provides a flexible, interactive annotation environment where each node can be linked out to its LocusLink description, and, in the case of gene expression networks, corresponding to Affymetrix probe set and chromosomal location information. GraphExplore can also integrate multiple networks, such as those generated under varying experimental conditions or those from different statistical models. Finally, GraphExplore produces a summary of the gene lists and statistical reports for a graph. It can export graphics in multiple output formats. GraphExplore is a full Java application and renders networks using Apache's Batik package (http://xml.apache.org/batik/); it can be launched from the main DIG query page and downloaded from the Duke website, together with full documentation and tutorial material.

The use of the various networks embedded in DIG to place a given gene into a functional context is illustrated in the Supplementary Material using an investigation of the Rb-E2F cell cycle pathway as an example.

Although other gene and pathway annotation tools have been developed in both academic and commercial settings, the DIG system provides an extremely useful combination of sources of gene annotation merged with the ability to explore complex networks that can extend the basic annotation analysis. The system has evolved out of project-based efforts to explore the significance of genes identified in gene expression profiling experiments and the need to have efficient and powerful mechanisms to place genes in a functional context. Such an integrative system facilitates the investigation of systems-level biological problems. The ability to understand these relationships, by identifying the relevant links that bridge distinct pathways, will be a critical next step in assimilating this knowledge into a unified, systems biology description of the activity within the cell.

REFERENCES

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Bhattacharjee,A. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Dobra,A. *et al.* (2003) Sparse graphical models for exploring gene expression data. *J. Mult. Anal.*, **90**, 196–212.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Huang,E. *et al.* (2003) Gene expression predictors of breast cancer outcomes. *Lancet*, **361**, 1590–1596.

Jenssen,T.-K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.

Jones,B., Carvalho,C., Dobra,A., Hans,C., Carter,C. and West,M. (2004) Experiments in stochastic computation for high dimensional graphical models. *Stat. Sci.*, in press.

Perou,C.M. *et al.* (2000) Molecular portraits of human breast tumors. *Nature*, **406**, 747–752.

Pittman,J. *et al.* (2004) Models for individualized prediction of disease outcomes based on multiple gene expression patterns and clinical data. *Proc. Natl Acad. Sci. USA*, **101**, 8431–8436.

Pomeroy,S.L. *et al.* (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–442.

Ramaswamy,S. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Rich,J. *et al.* (2005) Gene expression profiling and genetic markers in glioblastoma survival. *Cancer Res.*, **65**, 4051–4058.

Shipp,M.A. *et al.* (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, **8**, 68–74.

Singh,D. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 203–209.

van de Vijver,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

van'T Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

West,M. *et al.* (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.