

Internet tools for the Analysis of Gene Expression by Data Bases Integration

Angelo Facchiano ¹ and Alessandro Weisz ²

¹Istituto di Scienze dell'Alimentazione, CNR. Via Roma 52A/C - 83100 Avellino, ITALY. E-mail: angelo.facchiano@isa.av.cnr.it

²Dipartimento di Patologia Generale, Seconda Università di Napoli. Larghetto S. Aniello a Caponapoli, 2 - 80138 Napoli, ITALY

ABSTRACT

The DNA microarray technology provides experimental tools to evaluate the expression of thousands of genes at once, thus shortening dramatically the time needed for the experimental analysis. The main challenge, face to the very large data sets that are generated by these experiments, is represented by the need to evaluate them in detail, by comparing for example the simple numerical results obtained to the knowledge available on gene function and regulation. In fact, the simple information about the expression level of a specific gene is not relevant unless it can be evaluated taking into account also the role of each specific gene in the cell life. This step of the gene expression study is very time consuming due to the number of genes to be analyzed, as well as to the fragmentation of the information resources available. In this context, we are developing a package devoted to help this step in gene expression analyses. The package consists of PERL scripts and web interfaces, which allows to query the experimental data obtained by microarrays, integrate it with local databases and to explore the net searching for remote information useful for interpretation of the experimental results. Keyword or expression levels searches can be used to select interesting genes from the experimental data, saved as tables suitable for relational queries. Once the genes have been selected, their experimental data are formatted to be displayed in HTML with a simple color scheme, with links to public database entries concerning structural and functional gene-related information.

Information about the tool can be requested to angelo.facchiano@isa.av.cnr.it

INTRODUCTION

The last years were named "the genome era" because the challenge of the bio-medical research was represented by projects aimed at sequencing the whole genomes of many organisms. Now, while the human genome has been sequenced¹ and for other genomes this task will soon be completed, the "post-genome era" has started, in which the signals and the functions contained in the genome will have to be decoded. In fact, the knowledge of the sequence of a whole genome is not sufficient for a complete understanding of the many mechanisms governing the life of cells and organisms, for one reason because the regulation of gene expression is still unclear.

The new field of research named "functional genomics" is aimed to study gene expression by new technologies², mainly microarrays^{3,4}. The scientific community is working to create gene expression data repositories, as well as standards for DNA-array experiment annotation and data representation^{5,6}. These efforts are aimed mainly at allowing easy and reliable comparability of gene expression data from different sources, and interoperability of different gene expression databases and data analysis software. On the meantime, the Gene Ontology Consortium^{7,8} is working to create a controlled vocabulary to describe gene roles in any organism. In this context, task of the bioinformatics community is to create suitable tools to help researchers involved in gene expression studies^{9,10}. These tools should integrate and link the information resources (available as public data bases) to new experimental results obtained by laboratories involved in gene expression studies.

On this basis, we are working to develop a new visualising and analysing tool which can be used as a web server interface to query gene expression data bases. In the view of standardization of data bases, as well as of documents format, these tools are designed to work in agreement to the XHTML directives by the W3 Consortium^{11,12}.

SYSTEM AND METHODS

The tool consists of CGI query pages and PERL scripts which generate XHTML pages. The tools have been developed and tested on SGI Indy web site, but it should be easily installed and run under other Unix systems.

TOOL FEATURES

Figure 1 reports a schematic representation of the tool components. Three layers can be identified. The first one (on the left) is represented by the data bases themselves, both specific gene expression data and large public databases, as GenBank, GeneCards, OMIM, PubMed and so on. Local copies of gene expression databases are required, whereas the public databases can be in local or remote sites. The second layer (middle of the figure) is represented by the PERL scripts which take the queries, find the results into the gene expression data base, and create the output pages for the web server. The third layer (on the right) represents the user interfaces (input and output). Hyperlinks to public databases are created, according to user preferences, so that the output page can be optimized and linked to the local or nearest resources.

1. The first layer: the data bases

1.a) Data Bases of Gene Expression Data

Gene expression data are stored as flat files consisting of tables which report results and annotations. The first version of our tool was created to analyze results coming from a microarray lab. Now, this tool aims at being suitable for the standard formats used in the public repositories of gene expression data. However, these information must be integrated with more details, in order to allow links to other databases. As an example, gene expression data files from public repository do not report accession number reference to other databases, as GENBANK or others. Therefore, the gene expression data should be organized in more tables: the main table reports gene name and expression levels, whereas additional tables may report more information to link other resources. This structure should be suitable also for relational queries.

1.b) Data Bases of Sequences and Functional Data

It is very useful to link gene expression results to functional and structural information about the most interesting genes. In fact, similar expression levels of genes must be investigated in order to understand their possible relationships. However, the simple gene name reported in the results files is often not sufficient to this scope. Therefore, the visualization of results should report more information, or at least hyperlinks to databases where information can be found. Useful databases for this scope are GenBank, EMBL, UniGene, GeneCards, OMIM, PubMed, LocusLink. The current version of the tool uses remote hyperlinks to these resources. A more integrated access is planned.

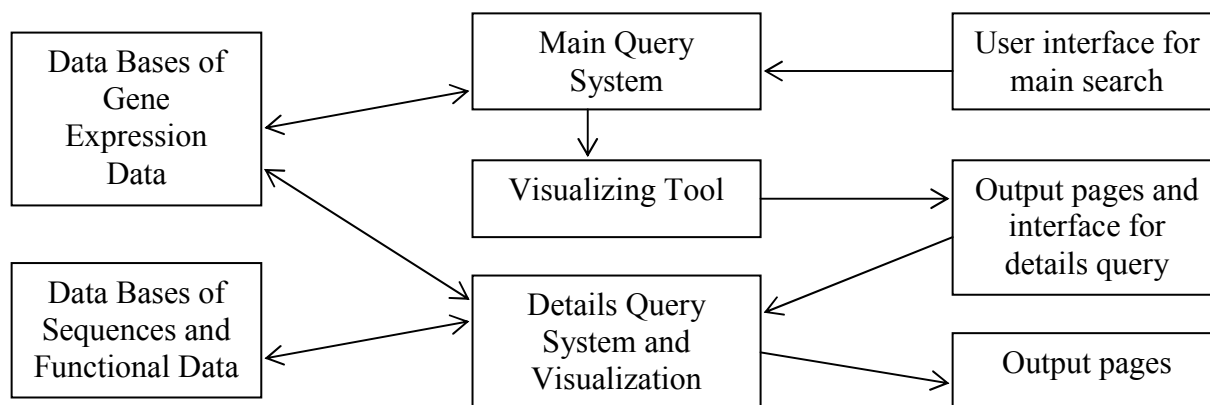


Figure 1. Schematization of the tool: the left of the figure represents data bases level, the middle represents PERL scripts level, the right represents the input - output level.

2. The second layer: PERL scripts

2.a) Main Query System

The core of the tool is a PERL script which is launched by the first query page, and search for the data base entries matching the query. The results are sent to the subsequent script for visualization. This script can be designed to exploit powerful query systems, depending on the system availability and the data base organization. The current version uses common Unix programs like "grep" and "awk" to extract data. Relational query systems can offer more powerful tools, and it is planned thier use to allow more complex searches.

2.b) Visualizing Tool

This tool creates one or more pages containing the gene expression data related to the genes selected by the Main Query System. The pages are created according to the XHTML format. Users can customize the output by setting the preferences in the main query page (see 3.a).

2.c) Details Query System and Visualization

This PERL script select the additional information, if available, for selected genes, and creates the output page to visualize fluorecence details as absolute fluorecence and spot images.

3. The third layer: the user interfaces

3.a) Main query page

This web interface allows to define the keywords to select genes, or the expression level required, as well as to set the format of the output pages. It is possible to customize the remote sites for data bases hyperlinks, the color scale, the number of genes to be shown for each page.

3.b) Main output pages and details query page

This page reports results concerning the selected genes, with links to local or remote resources. Numerical results are reported as a table, with colors related to expression levels, in order to obtain an immediate reading of the table. The implementation of graphs (for time-course experiments), which can be very useful to visual comparison of expression profiles, has also been planned. This page allows to select the most interesting genes and ask for more details.

3.c) Details output pages

This page reports the complete experimental information for the selected genes by showing the fluorecence absolute values, as well as the microarray spots, if available.

Future Perspectives

The tool is under continuous implementation. The most relevant improvement planned is the substitution of external links to public data bases by the content of such data bases. This task requires local copies of data bases, and suitable tools to extract the requested information related to the each gene under investigation. At the same time, the tool will be standardized in order to be suitable for new public gene expression data bases.

Availability

Information about the tool can be requested to angelo.facchiano@isa.av.cnr.it

Acknowledgements

Research supported by: AIRC (Grant 1998-2000), European Commission (Grant BMH4-CT98-3433), MURST (Cofin 1999), Seconda Università degli Studi di Napoli (Fondi per la ricerca 2000).

References

- 1 The human genome (2001). *Nature*, **409**, 813-958.
2. Brownstein M.J., Trent J.M., Boguski M.S. (1998) Functional genomics. *Trends Guide to Bioinformatics*, 27-29.
3. Watson A., Mazumder A., Stewart M, Balasubramanian S. (1998) Technology for microarray analysis of gene expression. *Current Opinion in Biotechnology*, **9**, 609-614.
4. Gaasterland T., Bekiranov S. (2000) Making the most of microarray data. *Nature Genetics*, **24**, 204-206.
5. <http://www.ncbi.nlm.nih.gov/geo/>
6. <http://www.mged.org/>
7. The Gene Ontology Consortium. <http://www.geneontology.org>
8. The Gene Ontology Consortium (2000) Gene Ontology: tool fo the unification of biology. *Nature Genetics*, **25**, 25-29.
9. Ermolaeva O., Rastogi M., Priutt K., Schuler G., Bittner M., Chen Y., Simon R., Meltzer P., Trent J.M., Boguski M.S. (1998) Data management and analysis for gene expression arrays. *Nature Genetics*, **20**, 19-23.
10. Bassett D.E., Eisen M.B., Boguski M.S. (1999) Gene expression informatics - it's all in yuor mine. *Nature Genetics Supplement*, **21**, 51-55.
11. XHTML 1.0: The Extensible HyperText Markup Language. <http://www.w3.org/TR/xhtml1>
12. XHTML 1.1 - Module-based XHTML. <http://www.w3.org/TR/xhtml11>