Simplifying Gene Expression Microarray Comparative Analysis.

Philip Church[1]**, Andrzej Goscinski**[1]**, Adam Wong**[1]**, and Christophe Lefevre**[2]
[1]School of Information Technology, Deakin University, Geelong, VIC, Australia
[2] Institute of Technology Research and Innovation, Deakin University, Geelong, VIC, Australia

Abstract - *Gene Expression Comparative Analysis allows bio-informatics researchers to discover the conserved or specific functional regulation of genes. This is achieved through comparisons between quantitative gene expression measurements obtained in different species on different platforms to address a particular biological system. Comparisons are made more difficult due to the need to map orthologous genes between species, pre-processing of data (normalization) and post-analysis (statistical and correlation analysis). In this paper we introduce a web-based software package called EXP-PAC which provides on line interfaces for database construction and query of data, and makes use of a high performance computing platform of computer clusters to run gene sequence mapping and normalization methods in parallel. Thus, EXP-PAC facilitates the integration of gene expression data for comparative analysis and the online sharing, retrieval and visualization of complex multi-specific and multi-platform gene expression results.*

Keywords: Gene Expression, Normalization, Clusters, Statistical Algorithms

1 Introduction

Comparative analysis is a fundamental tool in biology due to the influence of evolutionary and selective forces in shaping biological systems. Conservation among species greatly assists the detection and characterization of functional elements because important functional elements tend to be most conserved during evolution, whereas inter-species differences are likely indicators of biological adaptation. Comparative gene expression Analysis allows researchers to investigate the conserved or specific functional regulation of genes. Its basic principle is to group datasets based on gene evolutionary relatedness and isolate the components that behave in similar or different ways. Thus, comparing the regulation of genes in related organisms can assist the investigation of gene function. The microarray approach [1] is the most common method of collecting gene expression data currently being used in bioinformatics. More recently high throughput sequencing methodology is allowing an alternative approach for the estimation of gene expression. Data from microarray or sequencing experiments must be stored digitally using one of the many gene expression file formats before being analyzed using statistical algorithms and analysis. Normalization is a key part of gene expression microarray analysis since unnatural variations can be introduced during the data collection and digitization process. Thus, this data must typically be corrected, standardized and cross-referenced before being compared and analyzed.

Here, we present a web based package called EXP-PAC using the PHP/MySQL paradigm for the collaborative, integrative and comparative analysis of related gene sequences and gene expression experiments. The implementation also makes use of high performance computing to assist the integration, and analysis, of multiple gene expression datasets with common normalization methods and the inter-specific mapping of reference sequence datasets. Although the mapping of gene sequences between species has been performed and made available for a number of model organisms, for example in the Homologene database (http://www.ncbi.nlm.nih.gov/homologene), our package enables the rapid integration of sequence data collected from uncommon animal species, for which orthologous genome maps may not yet be referenced in public databases, and addresses the need of researchers working on a more diverse set of organisms or specific biological systems. For example we have developed an implementation of EXP-PAC dedicated to the integration and comparative analysis of gene expression during lactation in the mammalian lineage, which is accessible through the International Milk Genomics Consortium Web Portal (www.imgconsortium.org).

2 Gene expression comparative analysis

Gene expression comparative analysis is usually performed in the following three steps as illustrated in Fig. 1.

1. Data is collected in a wet-lab using a gene expression platform (cDNA, high throughput sequencing, Microarrays, etc.).
2. Collected data is converted to a digital format and any un-natural variation is removed.
3. Data analysis is used to group together similar datasets to locate components putatively responsible for biological functions.

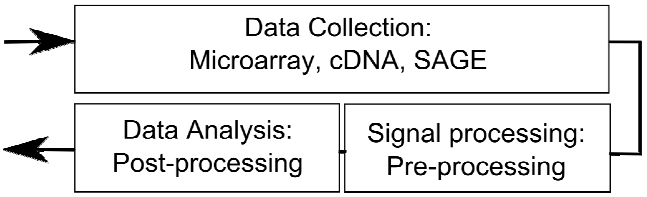

Fig. 1. The three stages of gene expression comparative analysis

2.1 Comparative transcriptome mapping

Before comparative gene expression analysis can proceed, the genes of one organism need to be cross-referenced to the related genes of another species. This is usually done through the identification of similarity by sequence similarity search algorithms such as BLAST [2]. Bi-directional reciprocal best hits often need to be identified and investigated for validation and identification of problematic gene family member assignment. Online access to sequences and maps greatly facilitates analysis of such gene family relationships and correct attribution of orthologous relationships for the construction of inter-genome maps. Although precompiled reference gene mapping data may be already available for a growing list of model organisms (Homologene), researchers working on non-model organism need to address the issue of cross-referencing genes. Our software package is built as an extension of an EST-PAC, a previously described package for the annotation of biological sequences [3]. Among other annotation tools, this package automates the management of sequence similarity search algorithms and the analysis of results through a web interface using a database and job management system. More recently we have implemented a new version allowing the use of high performance computing platform to optimize the performance of sequence similarity searches which is an important addition for the execution of multiple full genome searches required for the construction of inter-specific gene mapping as the execution time for bi-directional reciprocal mapping growth quadratically with the number of species. Once systematic sequence comparisons have been done, additional scripts can be deployed to compile cross-reference tables. The Unigene (http://www.ncbi.nlm.nih.gov/unigene) database maintains representative sequences for the genes of model organisms and, since many commercial gene expression platforms reference Unigene identifiers, we typical use these sequence references when available and build cross-references for other species or gene expression platforms using available cDNA libraries or transcript sequences predicted from related genome sequences.

3 EXP-PAC

EXP-PAC is a web-based system developed for the comparative analysis of gene expression. The EXP-PAC system combines the features of EST-PAC [3] with an on-line tool extension for the storage, analysis and visualization of gene expression data providing interfaces to facilitate SQL query based post-analysis of results (see Fig. 2.). EST-PAC is a sequence analysis framework, which provides online

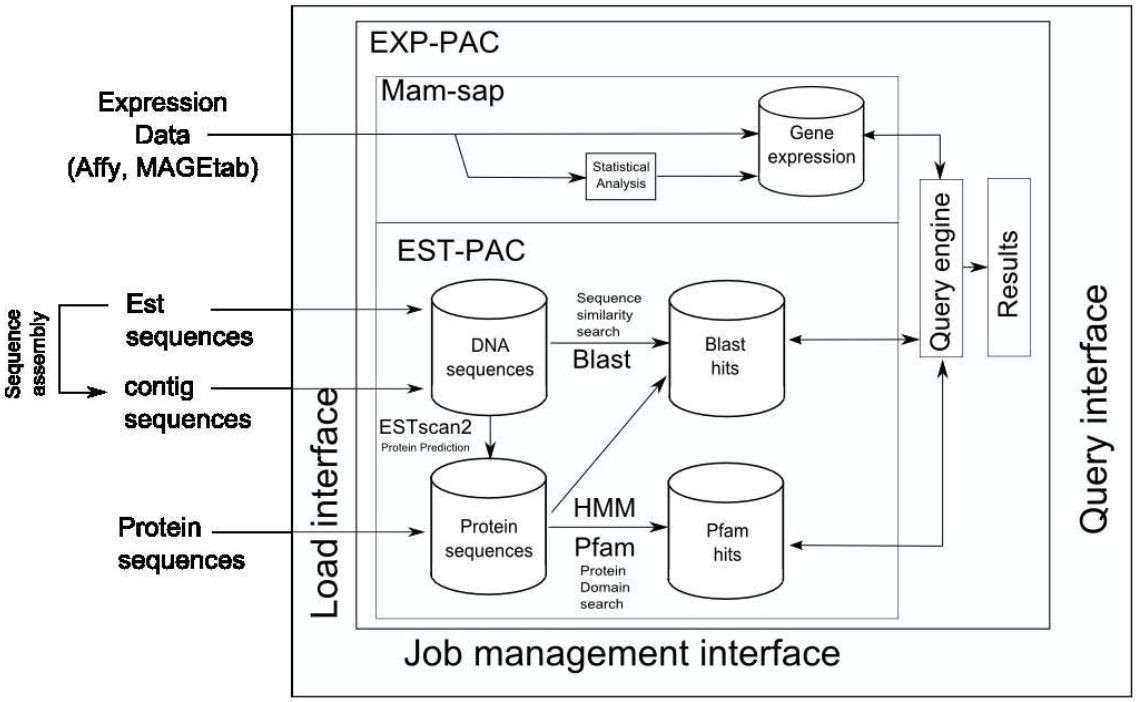

Fig. 2. The structure of the EXP-PAC system

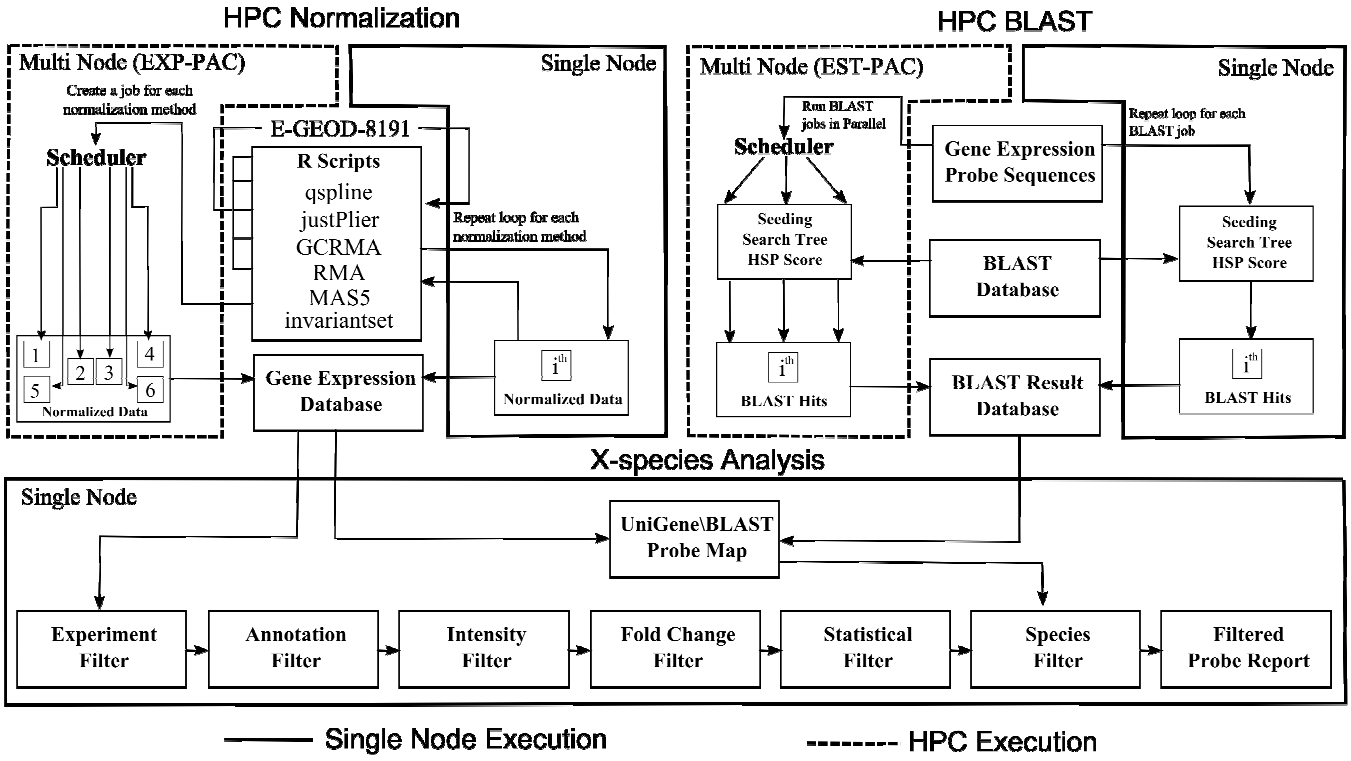

Fig. 3. EXP-PAC gene expression analysis workflow

interfaces for the storage of sequence data, the secure management of sequence annotation programs through embedded tools and, interfaces for the retrieval of sequence annotations. A new embedded high performance computing version of EST-PAC (EST-PAC^{HPC}) allows the cross referencing of transcriptome sequence catalogs through sequence similarity searches with the BAST program (companion paper in BIOCOMP11).

3.1 Annotation, data export and file sharing

The system allows the uploading of extendible gene annotation file formats associated with different gene expression platforms. This is necessary because gene annotation formats vary between gene expression platforms and data depositories. In addition, all data files uploaded in the system are archived and can be retrieved and downloaded from the interface, allowing data sharing and traceability.

3.2 Gene expression data upload and analysis

EXP-PAC provides users with the ability to upload a number of gene expression file formats (raw microarray data, SOFT [4], MAGE-tab [5], etc.) that may be available from download in gene expression databases [4, 6] or generated in the lab. Affymetrix microarray data files (also called CEL files) can be uploaded and automatically normalized with the R statistical scripting language using different established normalization methods for the Bioconductor package; including RMA [7], MAS5 [8], GCRMA [9] and PLIER [10] (Fig. 3.). EXP-PAC supports normalization through a distributed platform which uses the Sun Grid Engine [11] in order to speed up microarray data management and analysis for this common platform. By specifying the location of a bash script supported scheduler, normalization methods can be distributed over multiple nodes reducing the time taken for the normalization process. Other types of datasets can be normalized independently. Raw and normalized data can be uploaded and compared. Results from statistical analysis, obtained for example in the specialized Bioconductor package for R, can also be uploaded from tab-delimited files. Meta-data can be edited to group samples and adjust graphic display and color. Gene expression, associated gene annotation and statistical data can then be queried using an interface dynamically generated from the uploaded data. In addition, through creation of a sequence to probe ID map, it is possible for a user to perform comparisons on multiple species or experiments, retrieving the expression of likely orthologous genes (identified in the EST-PAC sequence similarity database) throughout a set of experiments in related species.

3.3 Query interface

EXP-PAC provides users with a web interface through which gene expression data can be queried (Fig 4.). A number of gene expression filtering methods are provided including; fold change, intensity levels, group average, probe ID and

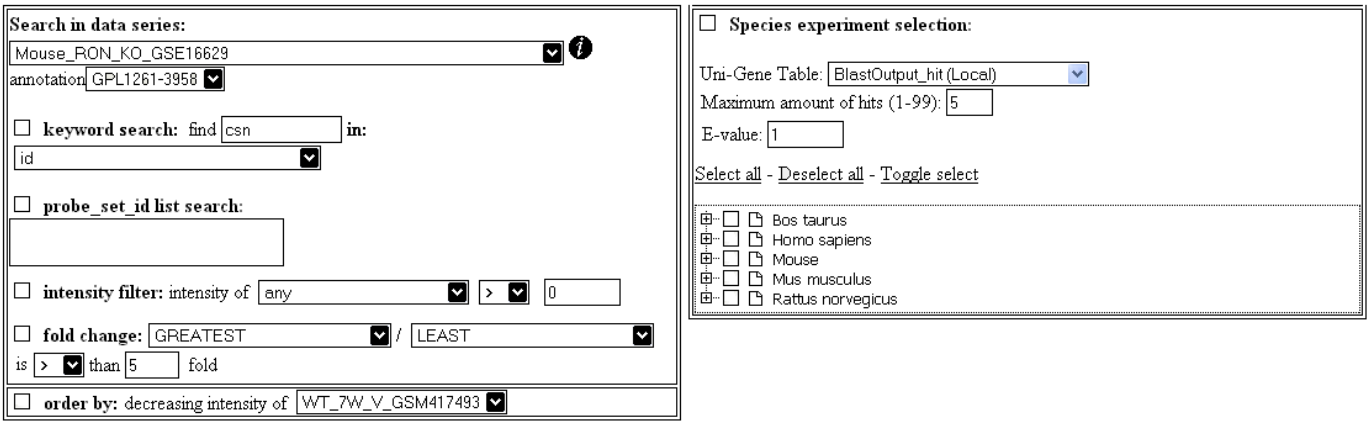

Fig. 4. EXP-PAC query interface

annotation. Returned probes can be ordered by selected probe intensity; displayed in descending order. Graphs are also produced to visualize the gene expression levels of each probe. A query builder tool allows users to create more complicated queries through generic interfaces that map to the SQL language. Using this tool, users create a database view by specifying tables and columns from list boxes. Created database views can be filtered using alphabetical and numeric values and operators. The results from created SQL queries can be saved or exported as a comma delimited text file. Visualization tools for investigating the distribution of gene expression data are provided to validate the normalization process. Users may also retrieve gene expression data across different species and experiments using pre-compiled reference maps of related probes and genes.

4 Conclusions

In this paper, we have presented an on-line framework for gene expression research. Compared to available gene expression software packages, EXP-PAC is unique in that it provides a method for the integration of cross-species gene expression experiments allowing comparative analysis and a method to perform high performance computing for reference sequence mapping and some common normalization methods. Most importantly, the EXP-PAC software package provides researchers with a simple way to manage and analyse gene expression and sequence data. SQL based analysis allow users to perform broad searches of stored datasets. In addition it is easy to integrate R scripts into the EXP-PAC system, allowing support for new and specialized methods and algorithms for gene expression or sequence analysis. Thus, EXP-PAC enables the development of analysis strategies integrating multiple experimental platforms in different species and provides an online workbench for comparative gene expression analysis.

5 References

[1] A. Brazma, *et al.*, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," *Nature Genetics*, vol. 29, pp. 365-371, 2001.

[2] S. F. Altschul, *et al.*, "Basic Local Alignment Search Tool," *Journal of Molecular Biology* vol. 215, p. 8, 1990.

[3] Y. Strahm, *et al.*, "EST-PAC a web package for EST annotation and protein sequence prediction," *Source Code for Biology and Medicine*, vol. 1, p. 2, 2006.

[4] T. Barrett, *et al.*, "NCBI GEO: mining millions of expression profiles--database and tools," *Nucl. Acids Res.*, vol. 33, pp. D562-566, January 1, 2005 2005.

[5] T. Rayner, *et al.*, "A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB," *BMC Bioinformatics*, vol. 7, p. 489, 2006.

[6] H. Parkinson, *et al.*, "ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression," *Nucleic Acids Res*, vol. 37, pp. D868-72, Jan 2009.

[7] R. A. Irizarry, *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostat*, vol. 4, pp. 249-264, April 1, 2003 2003.

[8] E. Hubbell, *et al.*, "Robust estimators for expression analysis," *Bioinformatics*, vol. 18, pp. 1585-1592, December 1, 2002 2002.

[9] Z. Wu, *et al.*, "A Model-Based Background Adjustment for Oligonucleotide Expression Arrays," *Journal of the American Statistical Association*, vol. 99, p. 909, 2004.

[10] I. Affymetrix. (2005, Technical note: guide to probe logarithmic intensity error (PLIER) estimation. Available: www.affymetrix.com/support/technical/technotes/plier_technote.pdf

[11] W. Gentzsch, "Sun Grid Engine: Towards Creating a Compute Power Grid," presented at the Proceedings of the 1st International Symposium on Cluster Computing and the Grid, 2001.