**Engene: the processing and exploratory analysis of gene expression data**

Jorge García de la Nava\(^1\), Daniel Franco Santaella\(^1\), Jesús Cuenca Alba\(^2\), José María Carazo\(^2\), Oswaldo Trelles\(^1\) and Alberto Pascual-Montano\(^2\),*

\(^1\) Computer Architecture Department, Universidad de Málaga, 29080 Málaga, Spain and \(^2\) Biocomputing Unit, National Center for Biotechnology, Campus Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain

Received on June 26, 2002; revised on September 24, 2002; accepted on November 2, 2002

**ABSTRACT**

Summary: **Engene** is a versatile, and platform-independent web tool for exploratory analysis of gene expression data that aims at storing, visualizing and processing large sets of gene expression patterns.

Availability: Access to **engene** is available on request at http://www.engene.cnb.uam.es

Contact: pascual@cnb.uam.es

**INTRODUCTION**

Gene expression monitoring technology allows to study the expression patterns of thousands of genes in a variety of experimental conditions and contexts and produces large numbers of data requiring automatic methods for their analysis. The analysis of these data is at the heart of bioinformatics concerns (see http://ihome.cuhk.edu.hk/~b400559/arraysoft.html for a review of currently available software).

The typical scenario for a bioinformatics research project has become a collaborative effort shared by several laboratories, producing, storing and analyzing common data in a coordinated way. To cope with this situation, bioinformatics web services seem a natural alternative to desktop-based applications since they provide universal access through a web interface from practically any web browser.

**Engene** is a platform-independent web tool for exploratory analysis of gene expression data that aims at storing, visualizing and processing large sets of gene expression patterns. **Engene** integrates a variety of analysis tools for visualizing, pre-processing and clustering expression profiles. Additionally, a Java suite for interactive analysis is also available. This tool enables the analysis of large sets of gene expression data in an easy and transparent way, allowing the study of the outcome of different pre-processing and clustering methods at the same time.

**IMPLEMENTATION**

**System architecture**

The core of **engene** is a C++ library of algorithms and data handling routines. On top of that library a group of applications that support the **engene** functionality is built. To increase the level of user friendliness and to allow the remote access to **Engene**, all the above applications have front-ends in a web server that are written in PHP to make the whole system more flexible and interactive (see Fig. 1).

The design of **engene** is suited for remote and multi-user operation. Since it follows the HTML standard, its portability is very high across web browsers. **Engene** mimics a directory tree with files that represent every type of supported data. There is an online help on every page for an instant short explanation of the options and algorithm parameters.

**Program overview**

**Engene** was developed to support different capabilities for exploratory data analysis. The data files are uploaded to the system as text files separated by tabs and, in addition to its own proprietary data format, **engene** supports different formats to be compatible with some of the freely available software packages. Once the data are loaded, its content is presented in a graphical representation of the measured signal intensities (Fig. 1) and some basic statistics of the data are also shown in the same window. At this point, there are different categories for the analysis that can be applied to the data set: pre-processing, clustering, mapping methods and statistical analysis. Pre-processing allows multiple operations to the data in order to accommodate it for further analysis. For example, filters to select genes of interest, normalization, logarithm transformation, treatment of missing data and others.

Once the data are properly transformed, several clustering and projection techniques can be applied. For example, classical techniques like hierarchical clustering.

\*To whom correspondence should be addressed.
(with different options for cluster agglomeration) and k-means. More sophisticated fuzzy clustering methods like fuzzy c-means and Fuzzy Kohonen Clustering Network (Bezdek, 1981) as well as kernel c-means (Pascual-Montano et al., 2001) are also included. The results of all these clustering methods can be graphically visualized in the web browser or can be interactively explored using a Java applet.

In addition, alternative projection techniques are also included. For example, linear mapping methods like PCA and non-linear techniques like Sammon mapping and Self-Organizing Maps (both classical and Batch SOM). These methods allow the projection of high dimensional gene expression data into a lower dimensionality space in such a way that they can be explored and visualized to detect their clustering structure. In this system, SOMs can be interactively explored via a Java applet, including visualizations like histogram, interneuron distance (U-matrix) and statistics of the clusters. Finally, new fuzzy and probabilistic self-organizing networks (FuzzySOM and KerDenSOM) were also integrated. These new SOMs are based on well defined mathematical cost functions expressing interesting properties of the projection process, obtaining not only a nonlinear projection but also a fuzzy membership and an estimated probability density function of the original gene expression data set (Pascual-Montano et al., 2001; Pascual-Marqui et al., 2001).

In summary, Engene provides an integrated environment for the exploration of gene expression data that includes not only the most used analysis algorithms in this field, but also some new and powerful clustering and mapping techniques. These analysis algorithms and their visualization tools make this system a good alternative to the currently existing software.

ACKNOWLEDGEMENTS
We acknowledge partial economic support by the Spanish CICYT, through grant Bio 2001-1237 and grant QLK3-2001-01473 of the EU.

REFERENCES