BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 17 2003, pages 2321–2322
DOI: 10.1093/bioinformatics/btg307

Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles

*Gabriel S. Eichler, Sui Huang and Donald E. Ingber**

Vascular Biology Program, Departments of Pathology and Surgery, Children's Hospital/Harvard Medical School, Boston, MA 02115, USA

Received on February 18, 2003; revised on April 11, 2003; accepted on May 27, 2003

ABSTRACT

Summary: Genome-wide expression profiles contain global patterns that evade visual detection in current gene clustering analysis. Here, a Gene Expression Dynamics Inspector (GEDI) is described that uses self-organizing maps to translate high-dimensional expression profiles of time courses or sample classes into animated, coherent and robust mosaics images. GEDI facilitates identification of interesting patterns of molecular activity simultaneously across gene, time and sample space without prior assumption of any structure in the data, and then permits the user to retrieve genes of interest. Important changes in genome-wide activities may be quickly identified based on 'Gestalt' recognition and hence, GEDI may be especially useful for non-specialist end users, such as physicians.

Availability: GEDI v1.0 is written in Matlab, and binary Matlab .dll files which require Matlab to run can be downloaded for free by academic institutions at http://www.chip.org/~ge/gedihome.html

Contact: donald.ingber@tch.harvard.edu

Supplementary information: http://www.chip.org/~ge/gedihome.html

Current analysis tools for high-dimensional gene expression profiles typically cluster similarly behaving genes or samples from different patients or tissue types. The main goal of these tools is to 'mine the data' for specific genes and determine their function. Common analysis algorithms, such as hierarchical clustering or self-organizing maps (SOMs) (Eisen *et al*., 1998; Tamayo *et al*., 1999), prejudge the results by forcing the data into a hierarchical organization (dendrograms) or a distinct number of clusters without any biological rationale. However, genes are components of a complex regulatory network which functions as an integrated dynamic entity (Huang, 2001). The existence of gene networks suggests that high-dimensional gene expression profiles may harbor global (genome-wide) patterns indicative of higher-order (emergent) phenotypic properties. A generic tool that provides a 'holistic' view of the transcriptome may therefore provide a means to identify system-wide patterns that are invisible in gene-centered clustering approaches.

Gene Expression Dynamics Inspector (GEDI) transforms dynamic or static, high-dimensional gene expression data into distinct two-dimensional (2D) color patterns. In doing so, GEDI brings abstract data to the realm of intuitive human 'Gestalt' perception that is often more powerful and versatile than existing pattern recognition algorithms. GEDI also displays the sample-to-sample distances indicating the speed at which the transcriptome travels through gene expression state space during a biological process.

At its core, GEDI uses an SOM (Kohonen, 1997) to reduce data dimensionality with respect to gene number and to create characteristic visual representations for each sample. However, unlike the conventional use of SOMs to classify genes or samples into discrete groups of potential biological significance, GEDI creates a global representation by using the SOM to project the genes onto a 2D mosaic where the tiles represent individual SOM clusters. Since the number of SOM clusters translates into the 'resolution' of the mosaic, we use SOMs with hundreds of 'miniclusters' (typically 1–9 genes) to create high-resolution mosaic pictures. The color of each tile of the mosaic is determined by the centroid value of that respective minicluster evaluated at each sample. Because SOMs place similar genes into the same neighborhood, coherent and robust pictures emerge that are characteristic of every sample. The spatial correspondence of the centroids is preserved because the mapping of genes to the tiles is invariant across the samples. For data sets composed of gene expression profiles from multiple time points, GEDI generates movies by animating the mosaic stack, providing an intuitive visualization of the time evolution of the transcriptome. (Examples can be seen at http://www.chip.org/~ge/gedihome.html). When static samples (e.g. different patients or tissue specimens) are compared, GEDI allows the user to browse quickly through the stack of mosaics to identify groups of similar patterns based on gestalt recognition, without prior definition of classes.

GEDI takes input data formatted in a tab-delimited ASCII file containing a $(N_P \times N_Q)$ matrix, where each of the N_Q columns q_{ji} represents a sample, and the N_P rows p_k represent the genes. The values of the matrix elements represent gene

*To whom correspondence should be addressed.

expression levels (e.g. log of expression change ratios). The set Q of the N_Q samples can be grouped in J classes, Q_J, i.e. $Q = \{Q_1, Q_2, \ldots, Q_J\}$ where each class may represent one of J time courses, tissues or experimental groups (if known a priori). Each class $Q_J = \{q_{J1}, q_{J2}, \ldots, q_{JT}\}$ contains T_J samples. Thus, an essential idea behind this analysis is that all the samples q of the entire set Q are concatenated column-wise in the $(N_P \times N_Q)$ matrix for training the SOM.

The output of the SOM is a mosaic of user defined size $m \times n$, consisting of $C = mn$ tiles which determine the 'pixel resolution' of the mosaic. The SOM maps each of the N_P genes of sample q to a tile c,which represents the centroid of a minicluster, $c = \{1, 2, \ldots, C\}$. Because of the concatenation of the samples in the input matrix, a gene x will be represented by the same tile (position) in each of the N_Q mosaics. This allows the direct comparison of multiple parallel time courses. Colors are assigned to each tile according to its centroid value. Stacks of mosaics of each sample class are displayed in the J windows. If the samples $q_{J1}, q_{J2}, \ldots, q_{JT}$ represent a time course, the mosaics are rearranged into their correct temporal order, interpolated and animated to produce a movie that runs in each of the J windows in synchrony or independently.

Figure 1A shows a simple time course experiment, i.e. $J = 1, Q = \{q_1, q_2, \ldots, q_t\}$ representing the time evolution of the gene expression profile during slime mold development (Van Driessche *et al.*, 2001). In Figure 1B, two time courses (i.e. two classes of samples) describing the development of male and female *Drosophila* are compared (Arbeitman *et al.*, 2002), thus, $J = 2, Q = \{Q_1; Q_2\} = \{q_{11}, q_{12}, \ldots q_{1T}; q_{21}, q_{22}, \ldots, q_{2T}\}$. In the slime mold example (Fig. 1A, top), the movie reveals a drastic change of the mosaic pattern between 6 and 8 h, indicating a 'jump' in gene expression state space, which is reflected in the peak of the inter-mosaic Euclidean distance (Fig. 1A, bottom) and coincides with a dramatic phenotypic transition from unicellular to multicellular state. Analysis of the mosaics from male and female *Drosophila* at different ages also immediately reveals a dramatic sex-specific phenotype: the mosaic patterns appear to be orthogonal (Fig. 1B).

GEDI allows the user to retrieve specific genes (and their annotations) associated with interesting regions of the mosaic patterns by clicking on a tile. This tool provides a mechanism to compare simultaneously gene activities across gene, time and sample space without prior assumption of any structure in the data. As bioinformatics and genomics begin to embrace a more holistic, systems view of biological processes, the visual power of GEDI also may be especially useful for non-specialist end users, such as biologists or physicians, who are more comfortable recognizing patterns (e.g. from a quick inspection of an X-ray) than a mathematical representation of the same pattern. GEDI v1.0 is written in Matlab R13 (Mathworks, Natick, MA) and distributed as a Matlab executable package for Windows only.

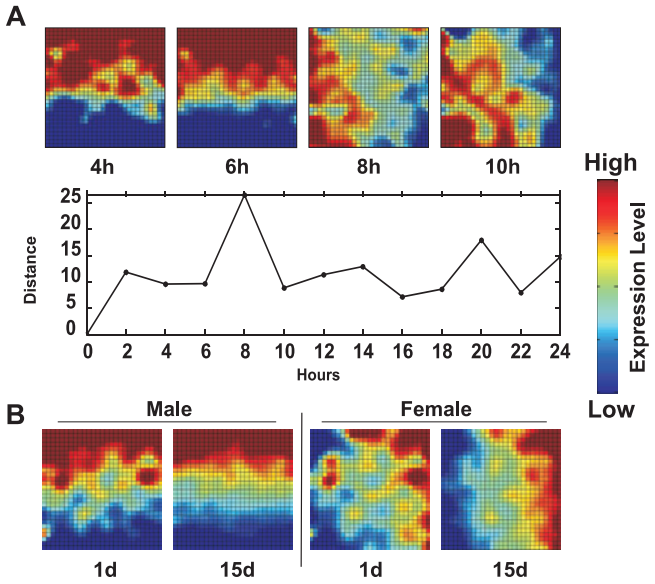

Fig. 1. Mosaics (individual frames from animations) and global metrics from published data sets produced by GEDI. (**A**) Top, mosaics visualizing time course of gene expression during *Dictyostelium discoideum* differentiation. Bottom, time course of inter-mosaic Euclidean distance (Distance). (**B**) Mosaics depicting male versus female expression patterns from *Drosophila melanogaster* at 1 and 15 day of development. The time to generate the mosaics scales with $O(K)$ where K is the user-defined quality of the SOM, typically, $K \sim N_p$.

ACKNOWLEDGEMENTS

This research was supported by grants from NIH to D.E.I. (NAG 2-1501), AFOSR to S.H. (F 49620-01-1-0564).

REFERENCES

Arbeitman,M.N., Furlong,E.E., Imam,F., Johnson,E., Null,B.H., Baker,B.S., Krasnow,M.A., Scott,M.P., Davis,R.W. and White,K.P. (2002) Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, **297**, 2270–2275.

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.

Huang,S. (2001) Genomics, complexity and drug discovery: insights from Boolean network models of cellular regulation. *Pharmacogenomics*, **2**, 203–222.

Kohonen,T. (1997) Self-organizing maps. *Springer Series in Information Sciences*, 2nd edn. Springer-Verlag, Berlin.

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.

Van Driessche,N., Shaw,C., Katoh,M., Morio,T., Sucgang,R., Ibarra,M., Kuwayama,H., Saito,T., Urushihara,H., Maeda,M. *et al.* (2002) A transcriptional profile of multicellular development in *Dictyostelium discoideum*. *Development*, **129**, 1543–1552.