Visualization of expression clusters using Sammon’s non-linear mapping

Rob M. Ewing\textsuperscript{1,*} and J. Michael Cherry\textsuperscript{2}

\textsuperscript{1}Carnegie Institution of Washington, Department of Plant Biology, 260 Panama Street, Stanford, CA 94305, USA and \textsuperscript{2}Genetics Department, School of Medicine, Stanford University, Stanford, CA 94305, USA

Received on January 24, 2001; revised on March 13, 2001; accepted on March 16, 2001

ABSTRACT

Summary: A method of exploratory analysis and visualization of multi-dimensional gene expression data using Sammon’s Non-Linear Mapping (NLM) is presented.

Availability: Scripts are available from the authors.

Contact: ewing@genome.stanford.edu

Large-scale gene expression data, such as generated by microarray experiments, is multi-dimensional; measurements are typically made of the expression of thousands of genes across multiple (10–100s) conditions. Analysis of large-scale gene expression data therefore requires visualization methods that are suitable for multi-dimensional data. Several approaches have been used to analyze multi-dimensional gene expression data, such as hierarchical clustering, Self-Organizing Maps (SOMs) and principal component analysis (reviewed in Sherlock, 2000). Common to most of these methods, is the identification of genes or experiments with similar expression profiles with subsequent clustering. Visualization then permits assessment of the degree of similarity within and between generated clusters.

We have used a technique for exploratory analysis of multi-dimensional data known as Sammon’s Non-Linear Mapping (NLM) (Sammon, 1969). NLM aims to map the input, high dimensional vectors to a lower dimensional space (typically two or three dimensions for visualization) whilst preserving any inherent structure in the data. Inter-object distances in the 2D or 3D space then approximate the inter-object distances in multi-dimensional space.

By way of example, we present data from the Arabidopsis Functional Genomics Consortium (AFGC) microarray project (http://afgc.stanford.edu), a project utilizing 11000-element cDNA microarrays to analyze gene expression in \textit{Arabidopsis thaliana}. Since a proportion of the array elements are redundant (i.e. represent the same gene), we sought to establish the degree of similarity between expression profiles of redundant array elements.

We used NLM as a first-pass method to screen these redundant sets of array elements and identify ‘outliers’ (array elements with expression profiles atypical of the set).

A data matrix consisting of expression measurements for 622 microarray elements (corresponding to the 100 most redundantly represented genes) measured across 28 microarray experiments was constructed (experiments were essentially a random subset from the AFGC project). The expression measurement used was the logarithm of the ratio of the intensity values of the two fluorescence channels. The 622\times28 expression data matrix was used to calculate the 622\times622 array element similarity matrix—using the linear correlation coefficient as the similarity metric. The input to the NLM algorithm is then the 622 vectors, each representing the set of correlation coefficients between a given array element and all other elements of the similarity matrix.

We used NLM as a first-pass method to screen these redundant sets of array elements and identify ‘outliers’ (array elements with expression profiles atypical of the set).

A data matrix consisting of expression measurements for 622 microarray elements (corresponding to the 100 most redundantly represented genes) measured across 28 microarray experiments was constructed (experiments were essentially a random subset from the AFGC project). The expression measurement used was the logarithm of the ratio of the intensity values of the two fluorescence channels. The 622\times28 expression data matrix was used to calculate the 622\times622 array element similarity matrix—using the linear correlation coefficient as the similarity metric. The input to the NLM algorithm is then the 622 vectors, each representing the set of correlation coefficients between a given array element and all other elements of the similarity matrix.

A subset of this data, 52 microarray elements corresponding to three different genes is shown in Figure 1. For comparison purposes, the data is displayed in three ways: as actual expression profiles (Figure 1a), as a dendrogram (Figure 1b) and as the equivalent 2D non-linear map (Figure 1c). Two outlying array elements corresponding to chlorophyll A–B binding protein (in pink, indicated by arrows) were identified. (Although as yet unconfirmed, the outliers are suspected to be truncated cDNA sequences corresponding primarily to 3’ UTR sequence.) We find that the NLM (panel 1c) allows clearest assessment of the ‘structure’ of the data, in this case identifying the two outlying elements, showing the similarity of the ‘red’ and ‘green’ profiles and showing the relative dissimilarity of the ‘blue’ profiles. Although we present a simple 2D representation here, we have also generated 3D NLMs, which might be further combined with more sophisticated visualization software.

Two additional strengths of visualizing expression clusters using NLM are as follows. A weakness of dendrograms is that arbitrary node orientation makes it difficult to assess the true distance between pairs of leaves (the ori-
Fig. 1. (a) Expression profiles of 52 microarray elements measured across 28 microarray experiments. Experiments are arranged across the x-axis (arbitrary order) and expression measurements (logarithm of the ratio of the two channels) on the y-axis. The 52 microarray elements represent three genes—ribulose bisphosphate carboxylase small subunit, chlorophyll A–B binding protein II precursor and hexameric polyubiquitin, colored in red, green and blue respectively. Expression profiles of two ‘outlying’ array elements for the gene represented in green are shown in pink and indicated by arrows (in (b) and (c) also). (b) Dendrogram showing relationships between expression profiles of the same data as in (a). Linear correlation coefficients were first calculated for all pairs of array elements (see text) and the UPGMA (Sokal and Michener, 1958) algorithm used to construct the tree. (c) Two-dimensional NLM representation of the same data as (a). The original NLM was calculated for the complete set of 622 array elements (see text) and the subset of 52 array elements (3 genes) as in (a) and (b) displayed. x and y are arbitrary units.

Acknowledgements
We thank Shu-hsing Wu and Katrina Ramonell for providing microarray data.

References