Biological Image Analysis via Matrix Approximation

Jieping Ye
Arizona State University, USA

Ravi Janardan
University of Minnesota, USA

Sudhir Kumar
Arizona State University, USA

INTRODUCTION

Understanding the roles of genes and their interactions is one of the central challenges in genome research. One popular approach is based on the analysis of microarray gene expression data (Golub *et al*., 1999; White, *et al*., 1999; Oshlack *et al*., 2007). By their very nature, these data often do not capture spatial patterns of individual gene expressions, which is accomplished by direct visualization of the presence or absence of gene products (mRNA or protein) (e.g., Tomancak *et al.*, 2002; Christiansen *et al.,* 2006). For instance, the gene expression pattern images of a *Drosophila melanogaster* embryo capture the spatial and temporal distribution of gene expression patterns at a given developmental stage (Bownes, 1975; Tsai *et al*., 1998; Myasnikova *et al*., 2002; Harmon *et al*., 2007). The identification of genes showing spatial overlaps in their expression patterns is fundamentally important to formulating and testing gene interaction hypotheses (Kumar *et al.*, 2002; Tomancak *et al.*, 2002; Gurunathan *et al.*, 2004; Peng & Myers, 2004; Pan *et al*., 2006).

Recent high-throughput experiments of *Drosophila* have produced over fifty thousand images (http://www.fruitfly.org/cgi-bin/ex/insitu.pl). It is thus desirable to design efficient computational approaches that can automatically retrieve images with overlapping expression patterns. There are two primary ways of accomplishing this task. In one approach, gene expression patterns are described using a controlled vocabulary, and images containing overlapping patterns are found based on the similarity of textual annotations. In the second approach, the most similar expression patterns are identified by a direct comparison of image content, emulating the visual inspection carried out by biologists [(Kumar *et al*., 2002); see also www.flyexpress.net]. The direct comparison of image content is expected to be complementary to, and more powerful than, the controlled vocabulary approach, because it is unlikely that all attributes of an expression pattern can be completely captured via textual descriptions. Hence, to facilitate the efficient and widespread use of such datasets, there is a significant need for sophisticated, high-performance, informatics-based solutions for the analysis of large collections of biological images.

BACKGROUND

The identification of overlapping expression patterns is critically dependent on a pre-defined pattern similarity between the standardized images. Quantifying pattern similarity requires deriving a vector of features that describes the image content (gene expression and localization patterns). We have previously derived a binary feature vector (BFV) in which a threshold value of intensity is used to decide the presence or absence of expression at each pixel coordinate, because our primary focus is to find image pairs with the highest spatial similarities (Kumar *et al.*, 2002; Gurunathan *et al*., 2004). This feature vector approach performs quite well for detecting overlapping expression patterns from early stage images. However, the BFV representation does not utilize the gradations in the intensity of gene expression because it gives the same weight to all pixels with greater intensity than the cut-off value. As a result, small regions without expression or with faint expression may be ignored, and areas containing mere noise may influence image similarity estimates. Pattern similarity based on the vector of pixel intensities (of expression) has been examined by Peng & Myers (2004), and their early experimental results appeared to be promising. Peng & Myers (2004) model each image using the Gaussian Mixture Model (GMM) (McLachlan& Peel, 2000), and they evaluate the similarity between images based on patterns captured by GMMs. However, this approach is computationally expensive.

In general, the number of features in the BFV representation is equal to the number of pixels in the image. This number is over 40,000 because the FlyExpress database currently scales all embryos to fit in a standardized size of 320×128 pixels (www.flyexpress.net). Analysis of such high-dimensional data typically takes the form of extracting correlations between data objects and discovering meaningful information and patterns in data. Analysis of data with continuous attributes (e.g., features based on pixel intensities) and with discrete attributes (e.g., binary feature vectors) pose different challenges.

Principal Component Analysis (PCA) is a popular approach for extracting low-dimensional patterns from high-dimensional, continuous-attribute data (Jolliffe, 1986; Pittelkow & Wilson, 2005). It has been successfully used in applications such as computer vision, image processing, and bioinformatics. However, PCA involves the expensive eigen-decomposition of matrices, which does not scale well to large databases. Furthermore, PCA works only on data in vector form, while the native form of an image is a matrix. We have recently developed an approach called "Generalized Low Rank Approximation of Matrices" (GLRAM) to overcome the limitations of PCA by working directly on data in matrix form; this has been shown to be effective for natural image data (Ye *et al.*, 2004; Ye, 2005).

Here, we propose expression similarity measures that are derived from the correlation information among all images in the database, which is an advancement over the previous efforts wherein image pairs were exclusively used for deriving such measures (Kumar *et al.*, 2002; Gurunathan *et al.*, 2004; Peng & Myers, 2004). In other words, in contrast to previous approaches, we attempt to derive data-dependent similarity measures in detecting expression pattern overlap. It is expected that data-dependent similarity measures will be more flexible in dealing with more complex expression patterns, such as those from the later developmental stages of embryogenesis.

MAIN FOCUS

We are given a collection of n gene expression pattern images $\{A_1, A_2, \cdots, A_n\} \in \Re^{r\times c}$, with r rows and c columns. GLRAM (Ye, 2005, Ye *et al.*, 2004) aims to extract low-dimensional patterns from the image dataset by applying two transformations $L \in \Re^{r\times u}$ and $R \in \Re^{c\times v}$ with orthonormal columns, that is, $L^TL = I_u$ and $R^TR = I_v$, where I_u and I_v are identity matrices of size u and v, respectively. Each image A_i is transformed to a low-dimensional matrix $M_i = L^T A_i R \in \Re^{u\times v}$, for $i = 1, ..., n$. Here, $u < r$ and $v < c$ are two pre-specified parameters.

In GLRAM, the optimal transformations L^* and R^* are determined by solving the following optimization problem:

$$\left(L^*, R^*\right) = \underset{L,R:L^TL=I_u, R^TR=I_v}{\arg\max} \sum_{i=1}^{n} \left\| L^T A_i R \right\|_F^2 .$$

Here, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix (Golub & Van Loan, 1996). To the best of our knowledge, there is no closed-form solution to the above maximization problem. However, if one of the two matrices L and R is given, the other one can be readily computed. More specifically, if L is given, the optimal R is given by the top eigenvectors of the matrix

$$\sum_{i=1}^{n} A_i^T L L^T A_i ,$$

while for a given R, the optimal L is given by the top eigenvectors of the matrix

$$\sum_{i=1}^{n} A_i R R^T A_i^T .$$

This results in an iterative procedure for computing L and R in GLRAM. For the given L and R, the low-dimensional matrix is given by $M_{i\,=} L^T A_i R$.

The dissimilarity between two expression patterns A_i and A_j is defined to be $\left\| M_i - M_j \right\|_F = \left\| L^T (A_i - A_j) R \right\|_F$. That is, GLRAM extracts the similarity between images through the transformations L and R. A key difference between the similarity computation based on the M_i's and the direct similarity computation based on the A_i's lies in the pattern extraction step involved in GLRAM. The columns of L and R form the basis