Citation:
Ye J, Janardan R & Kumar S (*2008*) Biological image analysis via matrix approximation. Encyclopedia of Data Warehousing and Mining (2nd Ed.) *edited by* John Wang, *Idea Group, Inc.*, Pennsylvania, 166-170.

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:

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

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

for expression pattern images, while M_i keeps the coefficients for the i-th image. Let L_j and R_k denote the j-th and k-th columns of L and R, respectively. Then, $L_j \cdot R_j \in \Re^{r \times c}$, for $j = 1, ..., u$ and $k = 1, ..., v$, forms the basis images. Note that the principal components in Principal Component Analysis (PCA) form the basis images, also called eigenfaces (Turk & Pentland, 1991) in face recognition.

We have conducted preliminary investigations on the use of GLRAM for expression pattern images from early developmental stages, and we have found that it performs quite well. Before GLRAM is applied, the mean is subtracted from all images. Using $u = 20$ and $v = 20$ on a set of *301* images from stage range *7--8*, the relative reconstruction error defined as

$$\sum_{i=1}^{n} \left\| A_i - LM_i R^T \right\|_F^2 \Big/ \sum_{i=1}^{n} \left\| A_i \right\|_F^2$$

is about *5.34%*. That is, even with a compression ratio as high as $320 \times 128/(20 \times 20) \approx 100$, the majority of the information (*94.66%*) in the original data is preserved. This implies that the intrinsic dimensionality of these embryo images from stage range *7-8* is small, even though their original dimensionality is large (about *40000*). Applying PCA with a similar compression ratio, we get a relative reconstruction error of about *30%*. Thus, by keeping the 2D structure of images, GLRAM is more effective in compression than PCA. The computational complexity of GLRAM is linear in terms of both the sample size and the data dimensionality, which is much lower than that of PCA. Thus, GLRAM scales to large-scale data sets. Wavelet transform (Averbuch *et al.*, 1996) is a commonly used scheme for image compression. Similar to the GLRAM algorithm, wavelets can be applied to images in matrix representation. A subtle but important difference is that wavelets mainly aim to compress and reconstruct a single image with a small cost of basis representations, which is extremely important for image transmission in computer networks. Conversely, GLRAM aims to compress a set of images by making use of the correlation information between images, which is important for pattern extraction and similarity-based pattern comparison.

FUTURE TRENDS

We have applied GLRAM for gene expression pattern image retrieval. Our preliminary experimental results show that GLRAM is able to extract biologically meaningful features and is competitive with previous approaches based on BFV, PCA, and GMM. However, the entries in the factorized matrices in GLRAM are allowed to have arbitrary signs, and there may be complex cancellations between positive and negative numbers, resulting in weak interpretability of the model. Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999) imposes the non-negativity constraint for each entry in the factorized matrices, and it extracts bases that correspond to intuitive notions of the parts of objects. A useful direction for further work is to develop non-negative GLRAM, which restricts the entries in the factorized matrices to be non-negative while keeping the matrix representation for the data as in GLRAM.

One common drawback of all methods discussed in this chapter is that, for a new query image, the pairwise similarities between the query image and all the images in the database need to be computed. This pairwise comparison is computationally prohibitive, especially for large image databases, and some ad hoc techniques, like pre-computing all pairwise similarities, are usually employed. Recall that in BFV (Kumar *et al.*, 2002; Gurunathan *et al.*, 2004), we derive a binary feature vector for each image, which indicates the presence or absence of expression at each pixel coordinate. This results in a binary data matrix where each row corresponds to a BFV representation of an image. Rank-one approximation of binary matrices has been previously applied for compression, clustering, and pattern extraction in high-dimensional binary data (Koyuturk *et al.*, 2005). It can also be applied to organize the data into a binary tree where all data are contained collectively in the leaves and each internal node represents a pattern that is shared by all data at this node (Koyuturk *et al.*, 2005). Another direction for future work is to apply binary matrix approximations to construct such a tree-structured representation for efficient image retrieval.

CONCLUSION

Identification of genes with overlapping patterns gives important clues about gene function and interaction. Recent high-throughput experiments have produced a large number of images. It is thus desirable to design computational approaches that can automatically retrieve images with overlapping expression patterns. The approach presented here (GLRAM) approximates a set of data matrices with matrices of low rank, thus avoiding the conversion of images into vectors. Experimental results on gene expression pattern images demonstrate its effectiveness in image compression and retrieval.

ACKNOWLEDGMENT

We thank Ms. Kristi Garboushian for editorial support. This research has been supported by grants from the National Institutes of Health and the National Science Foundation.

REFERENCES

Averbuch, A., Lazar, D., & Israeli, M. (1996). Image compression using wavelet transform and multiresolution decomposition. *IEEE Transactions on Image Processing, 5:1*, 4–15.

Bownes, M. (1975). A photographic study of development in the living embryo of *Drosophila melanogaster*. *Journal of Embryology and Experimental Morphology*, 33, 789–801.

Christiansen, J.H., Yang, Y., Venkataraman, S., Richardson, L., Stevenson, P., Burton, N., Baldock, R.A., & Davidson, D.R. (2006). EMAGE: a spatial database of gene expression patterns during mouse embryo development. *Nucleic Acids Research.* 34: D637.

Golub, G. H. & Van Loan, C. F. (1996). *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD, USA, 3rd edition.

Golub, T. *et al*. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286 (5439), 531-537.

Gurunathan, R., Emden, B. V., Panchanathan, S., & Kumar, S. (2004). Identifying spatially similar gene expression patterns in early stage fruit fly embryo images: binary feature versus invariant moment digital representations. *BMC Bioinformatics*, 5(202).

Harmon, C., Ahammad, P., Hammonds, A., Weiszmann, R., Celniker, S., Sastry, S., & Rubin, G. (2007). Comparative analysis of spatial patterns of gene expression in *Drosophila melanogaster* imaginal discs. In *Proceedings of the Eleventh International Conference on Research in Computational Molecular Biology*, 533-547.

Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

Kumar, S., Jayaraman, K., Panchanathan, S., Gurunathan, R., Marti-Subirana, A., & Newfeld, S. J. (2002). BEST: a novel computational approach for comparing gene expression patterns from early stages of *Drosophlia melanogaster* development. *Genetics*, 169, 2037–2047.

Koyuturk, M., Grama, A., and Ramakrishnan, M.-N. (2005). Compression, clustering, and pattern discovery in very high-dimensional discrete-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering*, 17(4), 447–461.

Lee, D.D. & Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788-791.

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models.* Wiley.

Myasnikova, E., Samsonova, A., Samsonova, M., & Reinitz, J. (2002). Support vector regression applied to the determination of the developmental age of a *Drosophila* embryo from its segmentation gene expression patterns. *Bioinformatics*, 18, S87–S95.

Oshlack, A., Chabot, A.E., Smyth, G.K., & Gilad, Y. (2007). Using DNA microarrays to study gene expression in closely related species. *Bioinformatics*, 23:1235-1242.

Pan, J., Guilherme, A., Balan, R., Xing, E. P., Traina, A. J. M., & Faloutsos, C. (2006).

Automatic mining of fruit fly embryo images. In *Proceedings of the Twelfth ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, 693–698.

Peng, H. & Myers, E. W. (2004). Comparing *in situ* mRNA expression patterns of *Drosophila* embryos. In *Proceedings of the Eighth International Conference on Research in Computational Molecular Biology*, 157–166.

Pittelkow, Y., & Wilson, S.R. (2005). Use of principal component analysis and the GE-biplot for the graphical exploration of gene expression data. *Biometrics*, 61(2):630-632.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S. E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S. E., & Rubin, G. M. (2002). Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biology*, 3(12).

Tsai, C. C., Kramer, S. G., & Gergen, J. P. (1998). Pair-rule gene *runt* restricts *orthodenticle* expression to the presumptive head of the *Drosophila* embryo. *Developmental Genetics*, 23(1), 35–44.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 71–86.

White, K.P., Rifkin, S.A., Hurban, P., & Hogness, D.S. (1999). Microarray Analysis of *Drosophila* Development During Metamorphosis. *Science*, 286(5447):2179-2184.

Ye, J. (2005). Generalized low rank approximations of matrices. *Machine Learning*, 61(1- 3):167-191.

Ye, J., Janardan, R., & Li, Q. (2004). GPCA: An efficient dimension reduction scheme for image compression and retrieval. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 354-363.

KEY TERMS

Compression Ratio: The ratio between the space needed to store the original data, and the space needed to store the compressed data.

Drosophila Melanogaster: A two-winged insect that belongs to the Order Diptera, the order of the flies. The species is commonly known as the fruit fly, and is one of the most widely used model organisms in biology, including studies in genetics, physiology and life history evolution.

Developmental Stage: A distinct phase in Embryogenesis, which is traditionally divided into a series of consecutive stages distinguished by morphological markers. In a high throughput experimental study, embryonic images have been grouped into six stage ranges, 1-3, 4-6, 7-8, 9-10, 11-12, and 13-16.

Dimensionality Reduction: The process of reducing the number of random variables under consideration, which can be divided into feature selection and feature extraction.

Embryogenesis: A process by which the embryo is formed and develops. It starts with the fertilization of the ovum, egg, which, after fertilization, is then called a zygote. The zygote undergoes rapid mitotic divisions, the formation of two exact genetic replicates of the original cell, with no significant growth (a process known as cleavage) and cellular differentiation, leading to development of an embryo.

Gene: A set of segments of nucleic acid that contains the information necessary to produce a functional RNA product in a controlled manner.

Gene Expression: A process by which a gene's DNA sequence is converted into functional proteins.