

CLUSTERING GENE EXPRESSION PATTERNS OF FLY EMBRYOS

Hanchuan Peng^{13*}, Fuhui Long³, Michael B. Eisen¹², and Eugene W. Myers¹³

¹ Genomics / Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA.

² Department of Molecular and Cell Biology, University of California, Berkeley, CA.

³ Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA.

ABSTRACT

The spatio-temporal patterning of gene expression in early embryos is an important source of information for understanding the functions of genes involved in development. Most analyses to date rely on biologists' visual inspection of microscope images, which for large-scale datasets becomes impractical and subjective. In this paper, we introduce a new method for clustering 2D images of gene expression patterns in *Drosophila melanogaster* (fruit fly) embryos. These patterns, typically generated from *in situ* hybridization of mRNA probes, reveal when, where and how abundantly a target gene is expressed. Our method involves two steps. First, we use an eigen-embryo model to reduce noise and generate feature vectors that form a better basis for capturing the salient aspects of quantized embryo images. Second, we cluster these feature vectors by an efficient minimum-spanning-tree partition algorithm. We investigate this approach on fly embryo datasets that span the entire course of embryogenesis. The experimental results show that our clustering algorithm produces superior pattern clusters. We also find previously unobserved clusters of genes that share biologically interesting patterns of gene-expression.

1. INTRODUCTION

RNA *in situ* hybridization provides a powerful way to visualize gene-expression patterns directly. Projects are underway to systematically collect RNA *in situ* expression patterns for a large number of genes during development in several organisms including nematodes [4], fruit flies [8][1] and mice [11].

The analysis of large-scale *in situ* datasets is by no means straightforward. Traditionally such images have been analyzed by direct inspection of microscope images, and several *in situ* databases record biologists' descriptions of expression patterns using a controlled vocabulary [8]. Automatic analyses are highly desired for annotation and many other applications. For example, identifying co-expressed genes from *in situ* expression data can dissect the gene expression networks that underlie development in multicellular organisms.

There has been little earlier work on image clustering to produce groups of genes with similar spatial-temporal patterns. This paper presents an efficient method for identifying clusters of similar *in situ* mRNA expression patterns. We develop the method in the context of the developing *Drosophila* embryo, using data from the Berkeley *Drosophila* Genome Projects *in situ* Database [1].

There are two major challenges:

- *Variation in embryo morphology, expression pattern staining and image orientation.* These factors often make the intrinsic cluster-structure obscure.
- *Large data dimensionality.* Often, each pixel is taken as one dimension, thus the dimensionality of an image sample is the total number of pixels, which is often very large (at

least around 100,000). High dimensional data not only presents a serious computational problem, but also contains lots of redundant information that needs to be removed in order to recover the intrinsic structures of the data distribution and its clusters.

To effectively cluster such large-scale noisy image data, there are two complementary issues to consider. First, what is a good representation of features in an *in situ* gene expression pattern? Second, how does one efficiently cluster the chosen feature representation? In this paper, we consider an eigen-embryo scheme to extract a feature vector for each embryo image by mapping the image to a low-dimensional eigen-image space (Fig. 1). Then, we introduce a very efficient tree-partition algorithm called MSTCUT to cluster these low-dimensional feature vectors of image patterns. We have performed comprehensive experiments on clustering several sets of *Drosophila* embryo patterns that span the entire course of embryogenesis. We show that for our data this approach is more efficient and accurate than several methods including spectral clustering and can find meaningful image clusters.

In the rest of the paper, we assume all the embryo images have been preprocessed using our earlier methods developed in [5]. The anterior-posterior axis of an embryo is horizontal (anterior always left, and dorsal always up).

2. EIGEN-EMBRYO FEATURES

The first step in our method is to generate feature vectors that characterize each image. Assume we have N images of *in situ* expression patterns I_1, I_2, \dots, I_N , each having M pixels. What is a good way to describe their features? One possible way as proposed in [5] is to detect prominent traits or Gaussian "blobs" in every image. Since different images can have different traits, this approach does not provide a canonical feature space against which the distribution of all image-patterns can be measured. Accordingly, clustering performance is limited.

An alternate approach is to decompose an image as a linear combination of a series of mutually orthogonal basis functions using the principal component analysis (PCA). This eigen-image analysis method, which was first applied to human-face recognition [9], uses the coordinates of an image in the eigen-image space as a feature vector that represents the original image. For our embryo images, this approach can be called the eigen-embryo analysis.

Fig. 1 illustrates the basic scheme of eigen-embryo analysis. In the pool of input images, two clusters are mixed together. Each cluster has a dark image, a bright image, and an image that is neither very dark nor very bright. Although visually we can see these two clusters of images have different expression patterns, it is not easy for a computer clustering program to separate them in the space of the raw image because the pixel intensity is the most pronounced feature in these 6 images, which obscures the cluster boundary. By projecting these images to the subspace of the first 3

principal eigenvectors (i.e. eigen-images), we represent each image using a feature vector w , which is the coordinate of image patterns in the 3D subspace. The cluster structure becomes apparent in the subspace, as indicated by small triangles and circles among which the cluster boundary can be easily drawn. Note that in the first PC, the dark curved structure (ventral nerve cord and brain) and light two-blob structure (guts) indicate the most prominent features to distinguish these 6 image-patterns.

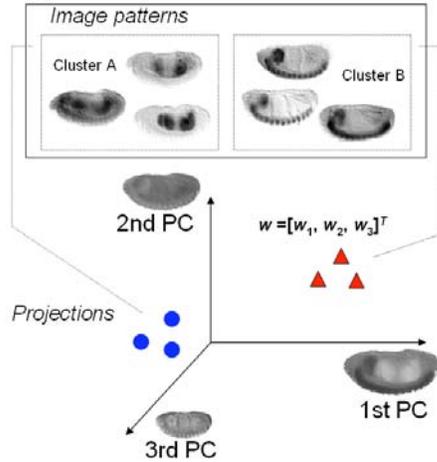


Fig. 1. Schematic illustration of the feature-vector generation and clustering using eigen-embryo analysis. For better visualization, we use the eigen-embryo sizes to indicate the ordering of principal components (PCs).

Mathematically, eigen-embryo analysis derives a short feature vector w to represent each image I , which is an $M \times 1$ vector. We first compute the centroid pattern of these N images, Θ . Each input image then differs from this centroid pattern by $X_k = I_k - \Theta$. Let $X = [X_1, \dots, X_N]$. Then the co-variance matrix is $C = XX^T/M$. The eigenvectors of the L largest eigenvalues of C form a subspace into which each image is projected by describing it as a weighted sum of these L eigen-embryos. This set of weights $[w_1, w_2, \dots, w_L]^T$ is a point in the L -dimensional eigen-embryo space and is called its feature vector. Denoting the i th eigenvector as v_i , the i th coordinate of the feature vector for image I is,

$$w_i = v_i^T (I - \Theta) \quad (1)$$

By describing each image in terms of the L most significant eigen-embryos the major variation of the data distribution is preserved, noise and redundancy are largely eliminated, and the distances between the feature vectors can be viewed as the distances between points in an L -dimensional space.

3. MSTCUT CLUSTERING

With each image now represented by a feature vector in the eigen-embryo subspace, we turn to describing a minimal spanning tree based clustering method called MSTCUT. We first construct a weighted, an undirected graph $G = (V, E)$, where V denotes a set of nodes and E denotes the set of edges between any pair of nodes. Each node $g \in V$ ($|V| = N$) is the feature vector of an image, and the edge weight s between a pair of nodes is the similarity of the respective image feature vectors. When the edge weight is 0, we

say two nodes are not connected. This graph can also be described as an $N \times N$ similarity matrix, S , of all the N nodes.

The image clustering problem can be defined as follows. Given N image data samples, we want to partition them into K pieces or clusters ($K < N$ and generally K is unknown) so that (1) each piece/cluster is a single connected component, (2) within each cluster the total similarity of data samples is maximized, and (3) across clusters the total similarity of data samples is minimized. Denote the similarity between the i th and j th samples as s_{ij} , where $1 \leq i, j \leq N$. Without loss of generality, we can assume the similarity score is between 0 and 1. Let $S = [s_{ij}]$ which is symmetric, A_1, A_2, \dots, A_K denote the K clusters, and S_{mn} ($1 \leq m, n \leq K$) denote the sum of the similarity matrix values of samples in A_m and A_n . Thus, the K -way clustering is to simultaneously optimize all conditions in Eq. (2), where K is unknown.

$$\begin{cases} \max S_{ii}, & i = 1, \dots, K \\ \min S_{ij}, & 1 \leq i, j \leq K, i \neq j \end{cases} \quad (2)$$

The criterion Eq. (2) is called **MinMaxPartition criterion**. One way to combine the optimization conditions in Eq. (2) is to optimize Eq. (3),

$$\min \sum_{i=1}^K \sum_{\substack{j=1 \\ j \neq i}}^K \left(\frac{S_{ij}}{S_{ii}} + \frac{S_{ij}}{S_{jj}} \right) \quad (3)$$

Generally, it is very difficult to solve the K -way partition exactly, mainly because it is highly combinatorial, the complexity of an exhaustive search is $O(K^N)$. An alternative method is to iteratively perform $K-1$ 2-way cuts. At each step we solve the much simpler problem of optimizing Eq. (4), which maximizes the intra-cluster similarity and minimizes the inter-cluster similarity simultaneously. In this case, a brute force search for the best partition has the complexity $O(2^N)$.

$$\min \frac{S_{12}}{S_{11}} + \frac{S_{12}}{S_{22}} \quad (4)$$

To be comprehensive we consider three widely used similarity scores, namely the L_1, L_2 Euclidean similarities and the correlation coefficient score. The L_1 and L_2 scores are computed from the respective Euclidean distance scores by applying a monotonically decreasing function to them. In this application, we use $s = \exp(-d)$, where d is the normalized Euclidean distance metric between two samples. This score tends to spread very similar points,

3.1 MSTCUT Clustering Algorithm

In the scenario of graph partition, a cluster of nodes is a connected component in the graph. Hence, for the fully connected graph G represented by the whole similarity matrix of all pairs of image samples, we can safely eliminate edges with the weakest similarity, while keeping the graph a connected component. Removing edges in increasing order of similarity or decreasing order of distance subject to not disconnecting the graph eventually leads to a minimum/maximum weight spanning tree (MST) depending on whether one considers distance or similarity, respectively, as the weight of an edge. An MST connects all nodes (image samples) in the graph, but has the minimum overall edge distance-scores. MST captures the basic cluster structures in the data, because the nodes that are more similar to each other are always connected in a shorter path in the tree.

An MST can be efficiently constructed using Prim's algorithm [2] that delivers a tree in $O(|E|+N\ln N)$ time when one uses a Fibonacci heap [2]. In our application the graph is not particularly sparse, i.e. $|E|$ is $O(N^2)$.

We propose a simple algorithm, called MSTCUT, to optimize the condition in Eq. (4). Because there are only $N-1$ edges in the tree graph and removal of one edge will bi-partition the graph, finding two clusters of nodes can be done by eliminating one edge in the tree. Since there are only $N-1$ different partition results, we can compare all of them to choose the best partition that minimizes Eq. (4) in $O(N^2)$ time.

How to generate K clusters? One way is to extend the MSTCUT to find K clusters simultaneously, by minimizing Eq. (3). This can be done in a brute force way to search all combinations of the $K-1$ edges dropped in the MST. The complexity is $O\left(\binom{N-1}{K-1} \cdot N\right)$.

Another more efficient way is to repeat the above bi-partition process on the sub-trees until K clusters are found. At each graph bi-partition step, we choose the cluster which has the smallest intra-cluster similarity for further partition. This cannot generate K clusters that are globally optimal, but the complexity is significantly reduced to $O(KN^2)$ time.

With the same MinMaxPartition Criterion, a very interesting method is called spectral clustering, which was developed in recent years to convert the combinatorial search problem in Eq. (4) as a linear search problem [7][3]. We note that in the limit case, we can apply the spectral clustering on the MST, which is the sparsest connected graph [6]. In this case, the partition result will be exactly the same as that generated by our MSTCUT method, because both spectral clustering and MSTCUT optimize the same function in Eq. (4). However, MSTCUT is faster, because it does not involve extra-computation of eigenvectors. We will show quantitative comparison of MSTCUT and spectral clustering in §4.

To evaluate the performance of clustering algorithm, we follow [6] and use the aggregated score of F -measures between the predicted clusters and the ground truth clusters. F -measure can be written in the form of precision P and recall R , which are widely used in information retrieval [10].

4. EXPERIMENT 1: COMPARISON OF DIFFERENT ALGORITHMS

Before we apply our approach to real examples in §5, we quantitatively compare our method against a few others.

For comparison, we selected several subsets of *mRNA* expression pattern images and determined the "ground truth" clusters of these images. Due to space limitation, here we only show results for one dataset, called P4Lateral, which corresponds to the lateral view of 167 embryos at the developmental stage 9-10 (i.e. phase 4 in [5], or about 4.8 ~ 6 hours after fruitfly egg hatching). In P4Lateral, there are three clusters manually determined by two human subjects. The consensus is used as the ground truth clusters. These three clusters include: (c1) 64 samples of gene expression patterns in the primordium of embryonic ectoderm regions (including procephalic, anterior, posterior, ventral, dorsal ectoderms), nervous system (like ventral nerve cord), and guts (like external foregut and inclusive hindgut). (c2) 44 samples of gene expression patterns only in the foregut and hindgut regions. (c3) 59 samples with artifact patterns that look like patterns in the ectoderm regions.

Results in Table 1 show that the introduction of eigen-embryos and MSTCUT improves the cluster-prediction. We see that for all three similarities L_1 , L_2 , and CC, the improvements of the eigen-image matching over the global matching, and those of MSTCUT over the spectral clustering, are significant. For example, the F -scores of our method are always higher than or around 0.9, whereas the F -scores of the comparing methods are around 0.7. Also note that for both clustering methods, the overall results for the eigen-image matching are better than those for global image matching. We also show the average result of random clustering. The average F -score of our method (above 0.9) is much higher than that of the random clustering method and very close to 1, indicating our results are very close to the manually generated "ground truth" and our method succeeds in finding the meaningful clusters in this dataset.

Table 1. F-measure scores of different clustering methods.

Clustering methods	Spectral clustering in [3]			MSTCUT		
	L_1	L_2	CC	L_1	L_2	CC
Similarity scores						
Global matching	0.69	0.69	0.57	0.69	0.70	0.74
Eigen-embryo Matching	0.74	0.69	0.67	0.92	0.93	0.89
Random clustering	0.346±0.002 (based on 20 trials)					

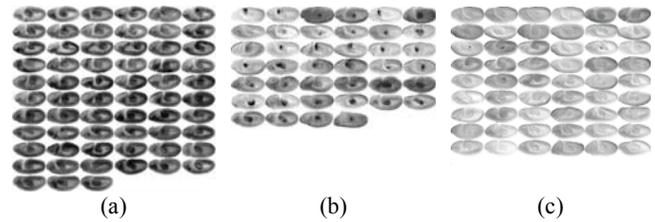


Fig. 2. Clustering results of P4Lateral dataset as three clusters in (a), (b) and (c).

5. EXPERIMENT 2: APPLICATION IN DETECTING CO-EXPRESSED GENES

It is known that co-expressed genes have similar spatial-temporal expression patterns over a range of embryo developmental stages. They might also be co-regulated in some modules in a genetic regulatory network. One way to detect co-expressed genes is to examine if several genes always have patterns that belong to the same cluster for multiple developmental stages. This is useful to infer if the respective genes share common regulators and how some genes are turned on/off under different conditions.

We design the following method to detect groups of genes that have similar expression patterns. The 16 developmental stages of *Drosophila* embryogenesis can be categorized as 6 phases, i.e. stages 1-3, 4-6, 7-8, 9-10, 11-12, and 13-16 [1], which coincide with major developmental transitions, e.g. gastrulation. Thus, we check the image clustering results phase-by-phase for every gene. The genes which have patterns in the same cluster of any specified phases are taken as a co-expressed gene-group in which genes share common spatial expression patterns during the entire course of embryogenesis. We have studied a set of 463 genes, with more than 4000 images. We have found many interesting clusters. Due

to space limitation, here we can only show one example of predicted *conditionally* co-expressed genes.

As shown in Fig. 3, three genes, Elongin-B (a transcriptional factor), Ngp (for GTP-binding) and CG17218, are found to share similar spatial patterns over the first 5 embryo development phases. We compared our prediction with the expert-annotations in the BDGP database [1], which have been listed besides the images in Fig. 3 (indicated by a "◆"). They also share a lot of common terms of image-ontology-annotation. However, for phase 6, the pattern of CG17218 is different from those of Elongin-B and Ngp. The respective expert-annotations for phase 6 are also different. This example shows that with our clustering method is able to detect these biologically interesting groups of gene clusters that might indicate the developmental-stage-specific co-expression/co-regulation relationships among genes.

As an ongoing work, we are currently investigating how to make use of these very useful gene phenotype clusters to detect sequence motifs of these genes.

REFERENCES

[1] BDGP: Berkeley Drosophila Genome Project (fruitfly.org).
 [2] Corman, T.H., Leiserson, C.E., and Rivest, R.L., Introduction to Algorithms, MIT, 2001.
 [3] Ding, C., He, X., Zha, H., Gu, M., and Simon, H., "A min-max cut algorithm for graph partitioning and data clustering," ICDM 2001, pp.107-114, 2001

[4] Hope, I.A., Albertson, D.G., ..., and Durbin, R. "The *C. elegans* expression pattern database: a beginning." Trends in Genetics, 12(9), pp. 370-371, 1996.
 [5] Peng, H.C., and Myers, E.W. "Comparing in situ mRNA expression patterns of Drosophila embryos," RECOMB 2004, pp.157-166, San Diego, USA, Mar, 2004.
 [6] Peng, H.C., He, X., and Long, F., "Automatic content extraction of filled form images based on clustering component block projection vectors," Proc. IS&T/SPIE 16th Annual Symp. of Electronic Imaging, Conf. on Document Recognition and Retrieval XI, pp.204-212, San Jose, CA, USA, Jan, 2004.
 [7] Shi, J., and Malik, J., "Normalized cuts and image segmentation," IEEE TPAMI, 22(8), 888-905, August 2000.
 [8] Tomancak, P., Beaton, A., ... Celniker, S.E., and Rubin, G.M., "Systematic determination of patterns of gene expression during Drosophila embryogenesis," Genome Biology, 3(12), 2002.
 [9] Turk, M., and Pentland, A., "Eigenfaces for recognition," Journal of Cognitive Neuroscience, 3(1):71--86, 1991.
 [10] van Rijsbergen, K. Information Retrieval, (2nd Ed.) Butterworths, London, 1979.
 [11] Visel A, Thaller C, and Eichele G., "GenePaint.org: an atlas of gene expression patterns in the mouse embryo," Nucleic Acids Res, 32, 2004. (genepaint.org)

Phase	Gene	Elongin-B (transcription elongation factor)	Ngp (GTP binding)	CG17218
1	stage1-3		maternal ◆	maternal ◆
			cellular blastoderm ◆	cellular blastoderm ◆
2	stage4-6		head mesoderm anlage ◆ posterior endoderm anlage ◆ trunk mesoderm anlage ◆	
			hindgut anlage	
			inclusive hindgut primordium ◆ salivary gland duct specific anlage salivary gland body specific anlage proventriculus primordium trunk mesoderm primordium ◆	
3	stage7-8		hindgut proper primordium ◆ posterior midgut primordium anterior midgut primordium salivary duct primordium salivary gland body primordium visceral muscle primordium	hindgut proper primordium ◆ tracheal primordium dorsal epidermis primordium foregut primordium
			embryonic/larval muscle system ◆ embryonic/larval somatic muscle dorsal prothoracic pharyngeal muscle embryonic/larval muscle system ◆	hindgut visceral branch tracheal system dorsal pouch
4	stage9-10			
5	stage11-12			
6	stage13-16			

Fig. 3. A predicted group of three *conditionally* co-expressed genes Elongin-B, Ngp, and CG17218, which have similar spatio-temporal patterns for the first 5 phases but not for the last phase (Elongin-B and Ngp still have similar patterns for phase 6). A "◆" is used to mark the common annotations.