Gene expression

Complex networks approach to gene expression driven phenotype imaging

L. Diambra and L. da F. Costa*

Institute of Physics at São Carlos, University of São Paulo, Caixa Postal: 369, CEP: 13560-970, São Carlos SP, Brazil

Received on May 18, 2005; revised on July 22, 2005; accepted on August 10, 2005

Advance Access publication August 16, 2005

ABSTRACT

Motivation: The need is to visualize and quantify gene expression spatial patterns. Because of their generality for representation of interaction among several elements, complex networks are used to measure the spatial interactions and adjacencies defined by gene expression patterns.

Results: Enhanced visualization of spatial interactions between elements where genes are expressed is possible, allowing the identification of structures which would go unnoticed by using conventional imaging. The quantification of the expression intensity in terms of the node degree and clustering coefficient allows the identification of different types of interactions, yielding insights about cell signaling and differentiation, and providing the basis for comparison and discrimination of the patterns along the developmental stages.

Availability: Supplementary Material, including visualizations as well as the basic routines for translating gene expression images into complex networks and obtaining node degree and clustering coefficient measurements, are provided.

Contact: luciano@if.sc.usp.br; diambra@univap.br

1 INTRODUCTION

The phenotype of an individual animal is a by-product of its genetic content as well as complex biophysical processes governing cell differentiation, organogenesis and animal development. By providing direct quantification of spatial gene expression, imaging techniques have established themselves as an important resource for quantifying and understanding animal development (Goldstein and Fyrborg, 1994; Müller and Newman, 2003).

Typically, the intensity of gene expression is represented by the gray-level values of the pixels in the imaged biological sample. Figure 1A illustrates an image of gene expression in Drosophila where each cell has been identified through image processing methods and used to calculate the respective intensity of gene expression (Kosma et al., 1999). A current problem regards the development of more effective resources capable of taking into account not only the expression intensity at the individual level, but also to relate such expression with the expression in its neighborhood. This endeavor is particularly relevant because gene expression-mediated pattern formation is characterized by spatial correlations and anticorrelations of cell activity achieved through intercellular signaling during the whole developmental process (Takaesu et al., 2002).

Established recently, complex networks research (Albert and Barabási, 2002) represents an interesting multidisciplinary area at the interface between graph theory and statistical physics, the latter being used in order to model and understand the dynamics of topological changes. Networks have been successfully applied to analysis and modeling of a large number of natural (e.g. Yeger-Lotem et al., 2004; Jeong et al., 2000) and human-related phenomena (e.g. Albert et al., 1999; Banavar et al., 1999). The most important property of a network, namely its connectivity, can be inferred in terms of topological measurements such as the node degree, corresponding to the number of edges connected to a node, and clustering coefficient, expressing the connectivity among the neighbors of a node (e.g. Albert and Barabási, 2002).

The current work proposes the application of geographical complex networks, where each node has a well-defined spatial position represented in terms of Cartesian coordinates, for visualization and integrated analysis of spatial gene expression derived from imaged gene expression intensities. Each expression element—e.g., a small volume of the organism or cell nucleus as in the example in Figure 1 and throughout this article, is represented by a node, while undirected edges are established between two nodes whenever the two following situations arise: (i) the nodes have similar expression intensity, and (ii) the nodes are no further apart from one another than a maximum distance. The former criterion is aimed at identifying intensity correlations between neighboring cells, a possible consequence of cell signaling and biochemical affinity. The latter criterion is imposed in order to constrain cell communication and emphasize the locality of gene expression. Although other criteria could be considered, e.g., aimed at studying anticorrelation (inhibition), the current work focuses on the two hypotheses above. As shown henceforth, such an approach provides the means not only to enhanced visualization of spatial relationships between expressing elements, but also to the effective quantification of such an activity.

This article starts by presenting the considered gene expression image database as well as the methodology suggested for translating gene expression patterns into complex networks. The achieved visualization and quantification features are then illustrated and discussed with respect to both synthetic (perturbed hexagonal lattices) and real gene expression patterns (Drosophila embryos), yielding a comprehensive quantification of the gene expression interactions.

2 MATERIALS AND METHODS

2.1 Experimental data

Gene expression pattern data from the FlyEx database for Drosophila embryos (http://flyex.ams.sunysb.edu/FlyEx/) are considered in this article. This database contains images of patterns at the translational level.
(i.e. protein levels) of several segmentation genes at different times during early development. The protein levels were measured by using fluorescently-tagged antibodies, as described in Kosma et al. (1998), which provide detailed information on gene expression at cellular resolution. Fluorescent images were obtained and used to get tables of gene intensity values for each nuclei. About 2500–3500 nuclei are described for each image. Each nucleus is characterized by an identification number, the position of its centroid, and the average fluorescence levels of three gene products. The centroid position over the anteroposterior (AP) axis is given by the percentage of embryo length, while the position over the dorso-ventral (DV) axis is expressed by the embryo width. The overall result is the conversion of an image to a set of numerical data which is then suitable for further processing. In this paper, we focus data related to the even-skipped (eve) gene.

All embryos under study belong to nuclear cleavage cycle 14 (Foe and Alberts, 1983). This cleavage cycle is particularly long (~50 min), and has been staged in 8 classes. All images were stained for eve, as well as two other proteins. In the present study we have considered the eve, Runt and Fushi-tarazu proteins which were stained simultaneously. In the early development of Drosophila, the segmentation process in the AP axis was driven by the gradient of bicoid protein (Bcd) (Driever and Nüsslein-Volhard, 1988a,b). Bcd is a transcriptional factor, and lies at the top of a cascade regulation. Both gap and pair-rule downstream segmentation genes are regulated by Bcd. In turn, gap genes also regulate pair-rule class genes, giving the typical narrow stripes of expression. In particular, the expression of the eve gene is mainly regulated by six (Bcd, Hb, Gt, Kr, Kri and Tll) upstream regulator factors, which form alternate domains of expression and repression along the AP axis. Thus, the nuclei in each domain share the same developmental history. In the present work, we take advantage of this fact to build networks that allow us to analyze the dynamic evolution of the segmentation process.

### 2.2 Complex networks approach

The gene expression data consists of the gene expression level at the nucleus, as well as the position of such nucleus (centroid), of the embryo cells. Our objective is to build a geographical complex network obtained by assigning the respective gene expression intensity to that node. The connection between a pair of nodes takes into account both the distance between pairs of nucleus and the protein level associated with such nodes. In particular, we have that any pair of nodes \( \{i, j\} \) is connected whenever the Euclidean distance \( \|\mathbf{P}_i - \mathbf{P}_j\| \) between them is less than \( D \), and if and only if the relative difference between the respective protein levels is smaller than \( \theta \). Mathematically:

\[
\text{edge}_{ij} = \begin{cases} 
1 & \text{if } \|\mathbf{P}_i - \mathbf{P}_j\| \leq D \text{ and } \frac{|\delta_i - \delta_j|}{\max(\delta_i, \delta_j)} \leq \theta \\
0 & \text{otherwise.}
\end{cases}
\]

![Fig. 1. Two typical networks (B and C) constructed from the same gene expression pattern (A) with different sets of parameters [(B): \( D = 35, l_{\text{min}} = 15, l_{\text{max}} = 35 \) and \( \theta = 0.11 \); (C): \( D = 35, l_{\text{min}} = 30, l_{\text{max}} = \infty \) and \( \theta = 0.20 \)]. The histograms in (D–G) characterize the obtained connectivity measurements. D and F depict the node degree histograms corresponding to the networks in B and C, respectively. E and G depict the cluster coefficient histograms obtained for the networks B and C, respectively. Note the different vertical axes scales.](https://academic.oup.com/bioinformatics/article-abstract/21/20/3846/202788)
number of edges attached to that node, the clustering coefficient by using the node degree is defined by

$$C_i = \frac{2e_i}{n_i(n_i - 1)}$$  \hspace{1cm} (2)

where $e_i$ is the number of edges among the set of nodes connected to $i$, and $n_i$ is the cardinality of this set. In order to characterize the network connectivity we consider the average value of $C_i$ over the whole network which is denoted by $\langle C \rangle$. Figures 2A and B illustrate a situation where $i$ has the same number of connections (i.e. same node degree) but whose connectivity between its neighbors is markedly different (i.e. different clustering coefficients). Contrariwise, the situation depicted in C shows the same clustering coefficient as in B but different node degrees. It is clear from these examples that the combined use of these two classical measurements are required in order to provide a more complete characterization of the interaction between the reference node and its neighbors as well as the surrounding connectivity. 

Figures 1B and C illustrate two examples of networks obtained from the same region of embryo gg17 (the gene expression image shown in A) considering different sets of parameters (Figure 1B: $D = 35$, $l_{\text{min}} = 15$, $l_{\text{max}} = 35$ and $\theta = 0.11$; Figure 1C: $D = 35$, $l_{\text{min}} = 30$, $l_{\text{max}} = \infty$ and $\theta = 0.20$). Therefore, the networks shown in Figures 1B and C reflect the similarity of gene expression intensities considering intermediate and high expression levels, respectively. A series of interesting visualization features have been allowed by the proposed network representation of gene expression, as can be identified by comparing the original intensity image in Figure 1A and the networks in B and C. One immediate advantage of the network representation is the clear indication of interactions between the expressing elements. Note also that the edges tend to align (along the tangent) with the level curves of the gene expression image, resulting orthogonal to the expression gradient. Combined with the intensity thresholding, such a property led to the approximated identification of the contour, shown in B, of the high gene expression region in the original image. In addition, more uniform gene expression regions tend to produce more dense and uniformly connected networks.

When the time interval between successive stages is short (as with the adopted data, characterized by ~6–7 min between frames), it is reasonable to assume that the changes in the gene expression pattern are driven by local cell signaling. Other similarity measures, such as temporal correlation, could be implemented as an alternative criterion for linking two nodes. However, often (as with current data) the nuclei positions are not maintained for any pair of embryo frames as a consequence of the use of different embryos at different stages during data acquisition.

Additional constraints can be imposed in order to restrict the analysis to high or low expression level domains. For example, in order to study domains with high (low) expression levels, nodes associated with protein levels lower (higher) than a threshold $l_{\text{min}}$ ($l_{\text{max}}$) are forbidden to connect to other nodes.

After building the network, we proceed to its respective characterization by using the node degree $k$ and the cluster coefficient $C$ (e.g. Albert and Barabási, 2002). While the degree of a specific node corresponds to the number of edges attached to that node, the clustering coefficient $C_i$ of a node $i$ is defined by

$$C_i = \frac{2e_i}{n_i(n_i - 1)}$$  \hspace{1cm} (2)

where $e_i$ is the number of edges among the set of nodes connected to $i$, and $n_i$ is the cardinality of this set. In order to characterize the network connectivity we consider the average value of $C_i$ over the whole network which is denoted by $\langle C \rangle$. Figures 2A and B illustrate a situation where $i$ has the same number of connections (i.e. same node degree) but whose connectivity between its neighbors is markedly different (i.e. different clustering coefficients). Contrariwise, the situation depicted in C shows the same clustering coefficient as in B but different node degrees. It is clear from these examples that the combined use of these two classical measurements are required in order to provide a more complete characterization of the interaction between the reference node and its neighbors as well as the surrounding connectivity.

Figures 1B and C illustrate two examples of networks obtained from the same region of embryo gg17 (the gene expression image shown in A) considering different sets of parameters (Figure 1B: $D = 35$, $l_{\text{min}} = 15$, $l_{\text{max}} = 35$ and $\theta = 0.11$; Figure 1C: $D = 35$, $l_{\text{min}} = 30$, $l_{\text{max}} = \infty$ and $\theta = 0.20$). Therefore, the networks shown in Figures 1B and C reflect the similarity of gene expression intensities considering intermediate and high expression levels, respectively. A series of interesting visualization features have been allowed by the proposed network representation of gene expression, as can be identified by comparing the original intensity image in Figure 1A and the networks in B and C. One immediate advantage of the network representation is the clear indication of interactions between the expressing elements. Note also that the edges tend to align (along the tangent) with the level curves of the gene expression image, resulting orthogonal to the expression gradient. Combined with the intensity thresholding, such a property led to the approximated identification of the contour, shown in B, of the high gene expression region in the original image. In addition, more uniform gene expression regions tend to produce more dense and uniformly connected networks.
Other prototypical expression patterns are shown in Figure 3, which were artificially constructed in order to emphasize how the proposed methodology reflects the spatial correlation of expression intensity. Figure 3A shows a case devoid of spatial correlation (random intensities), while the graphs in Figure 3B–C were obtained by imposing medium- and long-range spatial correlations. The highest node degrees were obtained for the more uniform situation in C, while the random intensity expressions in A yielded the smallest degree values, with an intermediate situation arising in B. The clustering coefficients tended to become higher and less disperse as one moves from less correlated (i.e. A) to more correlated expression intensities (i.e. C).

We conclude from the previous examples that more comprehensive characterizations of the complex networks derived from spatial gene expression images can be achieved by using both node degree and clustering coefficient. More specifically, we have: (i) the node degree tends to increase with the uniformity and correlation of the expression intensity, providing a direct indication of possible signaling between the cells; (ii) intermediate node degree values are obtained in situations characterized by a gradient of expression intensities such as in Figure 3B; (iii) higher clustering coefficients are obtained for more uniform expression patterns; (iv) the dispersion of cluster coefficient values is inversely related to the spatial uniformity of gene expression. Also, while the node degree reflects immediate signaling...
between cells, the clustering coefficient characterizes the gene expression context of a cell.

3 RESULTS AND DISCUSSION

Figure 4 depicts the complex networks obtained for Drosophila embryos at two different developmental stages from the spatial profile of eve protein (br2 and wn11 embryos from the FlyEx database). The anterior part of embryos points toward the bottom of the picture. These graphs considered \( D = 35 \) pixels, \( \theta = 0.10 \), \( l_{\text{max}} = 20 \) and \( l_{\text{max}} = \infty \) (i.e. those nodes corresponding to nuclei with protein levels below 20 were not attached to any node). At the first stage (embryo br2), the bulk of connections are mainly homogeneously distributed near the center of the embryo, reflecting the fact that protein levels are very homogeneous near the central domain. However, even at this stage, the edges at the anterior region are oriented perpendicularly to the AP axis. This edge organization is due to a eve protein gradient in that region. In contrast, the gradient at the posterior domain is weak, consequently the edges are oriented without preference. For stage 7, as the segmentation process evolves, community networks (i.e. clusters of more heavily connected nodes) arise. In this case, we observe an increase in the preference of edges for horizontal orientation following the typical stripes of eve pattern. It is interesting to note the enhanced visualization of gene expression interactions allowed by the complex network approach. Observe that several of the structures in the graph version of Figure 4-stage 1 would otherwise go unnoticed when observed directly from the gene intensity image (right-hand side of Figure 4-stage 1).

Figure 5 presents the histograms of node degree, cluster coefficients and edge orientations (angles respectively to the AP axis) for networks constructed from embryos at the same stages as shown in Figure 4. These histograms were computed over a set of 10 embryos. The error bars represent the standard deviation over this sample. The node degree evolves from a broad distribution with a long tail at the first stage to narrow distributions at stage 7, suggesting a decrease of immediate connectivity. At the first stage, the histogram of cluster coefficients almost corresponds to a Gaussian centered near 0.5. It evolves to a more intricate distribution with several narrow peaks, which have been verified to be pretty stable from stage 5 until the end of nuclear cleavage cycle 14 (data not shown). The edge orientation is, initially, mainly uniform with a weak preference for horizontal orientation. This preference becomes evident at stage 7. In order to characterize the histograms, we have also computed the entropies as well as the averages \( \pm \) standard deviations of the node degree and clustering coefficient values. The entropy is defined as

\[
S = -\sum p_i \log p_i,
\]

where \( p_i \) stands for the probability of event \( i \). Figure 6A depicts the entropy for the Drosophila eve gene expression data at successive developmental stages. Note that the entropy associated with the clustering coefficient distributions decreases along development as a consequence of the dispersed narrow peaks arising at developmental stage 4, while the entropy associated with the node degree decreases mainly because of the node degree concentration at low values observed at those same stages. It is worth noticing that different entropy evolutions have been obtained for other genes (for example fushi-tarazu, please refer to Figure S14 of the Supplementary Material). It is clear from these results that the node degree tended to decrease along development, indicating progressive spatial localization of gene expression interaction (Fig. 6B). At the same time, the clustering coefficient tended to increase steadily (Fig. 6C), implying that although each cell interacts with fewer neighbors, the interaction between such neighboring cells is more uniform and intense. In other words, the node degree and clustering coefficient evolution fully substantiate the progressive formation of the horizontal stripes along the Drosophila embryo, which are characterized by high uniformity.

4 CONCLUSIONS

We have described a new methodology for visualization and analysis of spatial gene expression patterns which involves transforming gene expression images into complex networks by using intensity similarity and distance constraints. The potential of the methodology has been illustrated with respect to both artificial patterns as well as Drosophila embryos at successive developmental stages. In addition to substantially enhancing the visualization of the spatial interaction among cells, which is potentially related to cell
signaling, the proposed approach allows the characterization of the putative interaction between cells. The biological interpretation of the node degree and clustering coefficient has been discussed with respect to a series of expression situations. The obtained results suggest that the proposed methodology, which can be applied to more general gene expression images, not necessarily involving the segmentation of the cells as in the adopted examples, can be widely applicable to the characterization of developmental dynamics, the identification of abnormalities, as well as comparative analysis between developmental dynamics in different animal species.

ACKNOWLEDGEMENTS
Luciano da F. Costa is grateful to FAPESP (proc. 99/12765-2), CNPq (proc. 308231/03-1) and Human Frontier Science Program (RGP 39/2002) for financial support. L. Diambra also thanks the Human Frontier Science Program for his post-doctoral grant.

Conflict of Interest: none declared.

REFERENCES