Published in final edited form as:

Mol Biosyst. 2013 July ; 9(7): 1926–1930. doi:10.1039/c3mb25475e.

A methodology to infer gene networks from spatial patterns of expression - An application to fluorescence *in situ* hybridization images

Monica Guimarães Campiteli^{a,*}, Cesar Henrique Comin^a, Luciano da Fontoura Costa^a, M Madan Babu^b, and Roberto Marcondes Cesar Jr^{c,d}

^aInstituto de Física de São Carlos - Universidade de São Paulo, Av. Trabalhador São-carlense, 400, São Carlos-SP, Brasil. Fax: 55 16 3372-2218; Tel: 55 16 3373-9758

^bMRC Laboratory of Molecular Biology, Hills Rd, Cambridge CB2 0QH, UK

^cBrazilian Bioethanol Science and Technology Laboratory, Brasil

^dInstituto de Matemática e Estatística - Universidade de São Paulo, Rua do Matão, 1010, São Paulo-SP, Brasil

Abstract

The proper functional development of a multicellular organism depends on an intricate network of interacting genes that are expressed in accurate temporal and spatial patterns across different tissues. Complex inhibitory and excitatory interactions among genes control the territorial differences that explain specialized cell fates, embryo polarization and tissues architecture in metazoan. Given the nature of the regulatory gene networks, similarity of expression patterns can identify genes and with similar roles. The inference and analysis of the gene interaction networks through complex networks tools can reveal important aspects of the biological system modeled. Here we suggest an image analysis pipeline to quantify co-localization patterns in *in situ* hybridization images of *Drosophila* embryos and, based on these patterns, infer gene networks. We analyze the spatial dispersion of the gene expression and show the gene interaction networks for different developmental stages. Our results suggest that the inference of developmental networks based on spatial expression data are biologically relevant and represents a potential tool for the understanding of animal development.

Introduction

The proper functional adult form of a multicellular organism is dependent on an intricate succession of events precisely coordinated in time and space¹. Gradients of transcription factors along the egg trigger the expression of a cascade of regulatory genes that will define the antero-posterior (AP) axis and tissue differentiation. Any mistakes in this regulatory network can lead to catastrophic consequences such as lethality. The study of developmental processes provides clues to the elucidation of developmental anomalies and evolutionary

problems. The fruit-fly *Drosophila melanogaster* is an important animal model given its facility of manipulation and the amount of respective data already available in the literature. Several databases dedicated to this organism are available on the internet ²⁻⁴.

High-throughput technologies have improved dramatically in the last decade and have enabled the access to hundreds or thousands of genes patterns of expression simultaneously. One of the main techniques are DNA microarrays, which allow the quantification of gene expression during specific events in genomic scale. Computational methods to analyze and compare such patterns have led to important findings ⁵⁻⁸. The result of a microarray experiment, however, is an average of the gene expression through a tissue or even a whole organism and the spatial information is completely lost. On the other hand, *in situ* hybridization (ISH) techniques ⁹ allow not only the measurement but also the localization of specific gene products in a tissue or organism as a whole. RNA ISH uses a RNA probe that hybridizes to a complementary mRNA sequence in the cytoplasm, revealing the gene's spatial patterns of expression. This yields a very powerful analysis integrating temporal and spatial information concerning gene regulatory events and the phenotypes associated. Development studies particularly benefits from such results ¹⁰⁻¹². The analysis of ISH images with the aim of elucidating genetic relations is gaining attention in the last years. Methods have been proposed to automatically extract and analyze gene expression patterns with the aim of gene clustering or automatic annotation ¹³⁻¹⁵.

Complex network approaches have proved to be a powerful tool to study biological systems due to their holistic nature ¹⁶. Networks are structures formed by nodes (that represent the system's elements) and edges, representing the interactions among them. This representation allows the analysis of the system as a whole and provides unique information about key genes or groups of genes and their roles in the cell ^{17,18}. We believe that the use of the spatial patterns of expression from RNA ISH data for the inference of gene interaction networks can considerably contribute to the understanding of biological systems. Thus, the current work proposes a method to infer gene interaction networks based on similarity of spatial expression patterns of *Drosophila* embryos. These patterns are represented as the projections of the gene expression densities along both antero-posterior and dorso-ventral axes and the networks are defined for four different developmental stages.

We present the pipeline developed to analyze ISH images and to infer the networks. We also present an application case-example to Fluorescent *in situ* hybridization (FISH) images and show results related to the spatial dispersion of the gene expression as well as the gene networks obtained for each developmental stages.

Accessing the development networks

Overview

Here we describe a methodology to process ISH images and to construct gene expression networks based on similarity of spatial expression patterns. It is commonly assumed that genes participating in a common pathway or regulated by the same transcription factors are likely to produce similar patterns of expression along time and this is well-explored in the literature ^{19,20}. Spatial expression patterns are formed in response to gradients of regulatory

genes along the embryo axis and interacting genes are commonly expressed in adjacent or superimposed areas ²¹.

Comparison of image patterns involves the design of methods to extract the images' main features while reducing dimensionality in order to make it computationally tractable. In this work we propose a methodology to process the embryos' images and to extract patterns for comparison. Relying on these patterns, it is possible to infer gene expression networks that allow a more robust analysis and understanding of the system involved. Figure 1 illustrates the pipeline. Each step is detailed in the respective following sections. In summary, we standardize the embryos orienting them along the horizontal axis, scaling and rotating in order to correctly locate the patterns along both the antero-posterior (AP) and dorso-ventral (DV) axes. This is finally represented by the sum of the pixel values over the two axes. The vectors represented by these projections (called feature vectors) are compared to define the edges of the networks.

Image pre-processing

In situ hybridization procedures commonly produce staining artifacts in the images, specially with fluorescent dyes. In order to deal with this we subject all the images to a pre-treatment that eliminates pixels below a given threshold of intensity. The threshold used represents 10% of the maximum pixel value.

ISH images contain mostly a single embryo fully captured in the picture. Nevertheless, a significant number of images include several embryos in the same picture or individuals not fully contained in the frame. In order to evaluate whether a particular image is suitable to our analysis, we determine the area of the object(s) contained in the picture using an erosion algorithm ²². A representative range of suitable areas is set and all objects above or below this range are deleted from the picture as well as all the objects touching the edges of the picture. This step is intended to assure the analysis of integral individual embryos.

The correct location of the expression patterns according to the head and tail is one of the main points for the analysis of correlated genes. Thus one of the main steps in our analysis is the correct standardization of the images. Patterns along the antero-posterior (AP) axis contribute most in our analysis, thus the head-tail orientation was performed manually to prevent any misclassifications to propagate through the pipeline. In ¹³, a heuristic analysis was performed to determine AP orientation based on the maximization of some similarity measure between images representing the same gene and same situation. However this method is dependent on the existence of such pairs of images and on the previous annotation by the image bank. In this work, we did not perform dorso-ventral orientation. Although this seem to represent a source of error, for the data set used in this work, we found that the dorso-ventral information positively contributed to our results.

The embryos in those images can appear rotated by any angle around the horizontal axis. Embryos were projected into the principal axis using Principal Component Analysis (PCA) ²². PCA is a linear orthogonal transformation that projects data over a new coordinate system assuring that the largest variance occurs along the first axis. Following projection,

the figures are cut tangentially to the embryos edges and the result is scaled to a standard length.

Image processing

The net result of the procedure described above is a set of images with unified length and width showing the patterns stained by the ISH dye. The last processing step is to sum pixels along the AP and the DV axes according to the example in Figure 1, in order to obtain an estimate of the patterns of expression along the embryos' two main axes. For fluorescent ISH (FISH) the mRNA molecules are stained by a green fluorescent dye while the nuclei of cells are stained red. In order to eliminate possible contributions due to the shape of the embryo, vectors obtained from green and red layers are subtracted from each other as follows:

$$v_i = G_i - R_i, \quad (1)$$

R_i being the number of non-zero pixels along row/column i in the red layer and G_i is the sum of green pixel values along row/column i . Each vector (G_i and R_i) is previously standardized between 0 and 1. While the red pixels are considered only to account for the embryo's shape, the intensity of the green fluorescence is proportional to the transcripts concentration and must, therefore, be accounted. On the other hand, considering the intensities of red pixels, adds a considerable amount of noise to the equation. This procedure is performed independently for the AP and the DV axes. The final feature vectors are obtained from the concatenation of the vectors extracted from each axis. In this work, the feature vectors are represented in 700 dimensions, corresponding to 500 AP columns and 200 DV rows.

Network inference

The steps aforementioned are meant to adequately reproduce the expression patterns along the embryo's axes. The next important task is to conceive an efficient measure of similarity for the specific context of the obtained vectors. Among the most commonly used measures to evaluate similarity between gene expression patterns are Pearson linear correlation¹⁸ and Mutual Information (MI)²³. For the sake of a more intuitive interpretation and a low computational cost, in this work, we chose to use Pearson correlation, defined as:

$$r_{xy} = \frac{1}{N} \sum_i \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right), \quad (2)$$

where x_i is the i^{th} observation of gene x and $\bar{x}$ is the mean value of x .

The result of the calculation is a matrix where each element r_{xy} represents the correlation between patterns of expression of genes x and y (according to our assumptions). In order to define the network, it is necessary to set a cutoff value r_0 above which a pair of nodes is considered connected. For this work, we defined $r > 0.75$ as a suitable cut-off value. The distributions of correlation coefficients are given in the Supplementary Material along with the distribution of correlations obtained for a random ensemble of shuffled vectors (p – value $< 10^{-5}$).

Drosophila FISH images

We applied the proposed method to an image set down-loaded from <http://fly-fish.ccbr.utoronto.ca>³. This website, maintained by the University of Toronto, hosts drosophila FISH images for 4030 genes classified in one of five stages of development. The pictures are taken in many angles and there is no annotation concerning the view. FISH images are stained for specific mRNAs with a green fluorescent probe. Nuclei are stained in red so it is possible to identify the boundaries of the embryo.

For this work we downloaded a set of about 33000 images concerning 2522 genes (corresponding to a previous update of the database). However, most of this set is composed of ubiquitous patterns, i.e. images homogeneously stained, lacking a clear pattern of expression. These images contribute nothing to the analysis and can interfere with the results. Thus, we have manually excluded all images with homogeneous staining. The remaining set is greatly reduced and the number of genes for each stage is listed in Table 1.

This set generated five networks corresponding to each development stage. The first stage was not included in the analysis due to its very small number of genes. The results presented subsequently concern the networks from the second to the fifth stages.

Data analysis

In order to have an idea of the overall dispersion of the feature vectors of the individuals, we performed a PCA into two variables for each of the stages. The weights of the two first principal variables for stages 2 to 5 are presented in Figure 2. The percentage of explained variance varies from 40% to 50%, indicating that the dispersion of the 700 original features can be concentrated at a surprising rate along the first two principal variables. In Figure 2 one can notice that the contributions of the different portions of the embryo (in absolute value of the weights) are very similar for all the four stages in the first variable. Regarding the first principal variable, it can be inferred that the anterior portion of the embryo (region **A**) has a particularly important contribution to the variance of gene expression, followed by the posterior extremity (region **C**). On the other hand, the second principal variable seem to discriminate the first two stages from the fourth and fifth. In stages 2 and 3, the posterior end seems to represent the greatest importance, alternating to the head portion in stages 4 and 5. Concerning the dorso-ventral-projection (region **D**), we can also see a distinction between the first two stages and the later ones. DV axis seems to have a greater contribution to the overall variance of gene expression in stages 2 and 3.

Developmental networks

The networks obtained for stages 2 to 5 are represented in Figures 3(a) to 3(d).

The obtained networks are composed of a number of disconnected subnetworks. We treated each subnetwork as a community and were able to reasonably relate most of them to a defined pattern of spatial expression or embryonary region (to facilitate the analysis, the subnetwork differentially colored in Stage 4 was manually splitted into three subgroups based on the low density of connections among their nodes). We compared the groups found with our method to the annotations available in the Berkeley Drosophila Genome Project

(BDGP) in situ homepage <http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl> where the images are annotated with a controlled vocabulary for developmental anatomy of the fly embryogenesis. In order to establish statistical significance, we run a hypergeometric test to verify enrichment for the annotations ($p < 0.05$). The groups showing enrichment are highlighted in different colors identified by a summarizing term. The p-values obtained for each group as well as the original annotations are given in the Supplementary Material.

The patterning process during the early stages of development in *Drosophila* is very well understood²¹. During the first stages, organ differentiation or cell migration are not expected. The embryo contains a layer of superficial nuclei called blastoderm and the yolk. The antero-posterior and dorso-ventral axes are established. Pole cells, the precursors of the germ line, segregate from somatic cells at the posterior region of the embryo. After a series of nuclei divisions, the embryo cellularize and initiates gastrulation. During this phase, massive movements of cells will give rise to the tissue layers from which tissue differentiation occur - endoderm, mesoderm and ectoderm. By this time, the fates have already been established for many cells²⁴.

In Figures 3(a) and 3(b), the predominant events are related to the establishment of the main axes and the determination of the tissue and organs primordia. This is also reflected by the results in Figures 2(a) and 2(b).

Among the most important genes coordinating AP determination are *Bicoid* (*bcd*) and *hunchback* (*hb*) that are responsible for the patterning of the anterior parts of the embryo - head and thorax. These genes work in contrast with two others - *nanos* (*nos*) and *caudal* (*cad*), that regulate the establishment of posterior structures. Those are called Maternal effect genes and work in tight coordination. *Bicoid* is present in the unfertilized egg and concentrates at the anterior tip of the embryo while *nanos* concentrate at the opposite end. The presence of *bcd* inhibits the expression of *cad* and *nos* blocks the expression of *hb* at the posterior region of the embryo. The expression of these genes regulate the expression of a group of segmentation genes that will eventually determine the larval segments.

Two significant enriched subnetworks in the second development stage are composed of genes expressed along with *bcd* at the anterior portion of the embryo and *hb* at the procephalic region, respectively. The genes *nos* and *cad* are not present in the database.

In a later embryogenesis phase, during segmentation, the homeotic selector genes play a major role providing positional information along the antero-posterior axis and determining the differentiation of the thoracic segments. *Antennapedia* (*Antp*), *Ultrabithorax* (*Ubx*) and *teashirt* (*tsh*) are known to interact in different ways to achieve gut differentiation, dorsal vessel polarity and muscle cell fate by acting on the mesoderm cells²¹. Consistently, these genes appear together in the network corresponding to stage 4. *Ubx* also appears in our analysis set in stage 5, related to posterior gut differentiation (amnioserosa).

It is worth mentioning that, although our method does group together known interacting genes, we were not able to show statistically significant results for a pair-wise analysis of the networks, comparing to reported genetic or physical gene interactions. Genetic

interactions leading to embryogenesis are highly complex and we believe that incorporation of temporal information is essential to gene-to-gene analysis.

Conclusion

With the advances in high-throughput methods, the development of techniques to automatically analyze and extract information from a large amount of data has become imperative. This paper proposes a simple methodology to analyze ISH images and to extract the gene patterns of expression along the AP and DV axes with low computational cost and easy implementation. This method has been illustrated with respect to fluorescent *in-situ* hybridization obtained from the FISH database³. A principal component analysis showed that the AP projection accounts for a large percentage of the gene density variation along space, with the head region being more prominent in this respect. The DV projections, however, showed different contributions for the early stages in comparison to the later ones regarding the spatial variance of gene expression.

We obtained networks with distinct groups of genes that demonstrated a high biological coherence. The method can be used as a tool to explore new genes' functions and relations. We stress that the process to obtain the networks is largely unsupervised, paving the way for applications of the reported methodology to other staining techniques as well as other organisms.

As the image techniques improve and more data become available, more information can be incorporated to the method, for instance, expression along time, allowing a more thorough analysis of the results.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Dr. Zila L. P. Simoes for a valuable support in the analysis and discussions of the results. The authors also thank Dr. Rita Zorzenon for discussions and suggestion regarding normalization. RMC would like to thank FAPESP and CNPq for financial support. LFC acknowledge support from the Human Frontier Science Program RGP39/2002, FAPESP (procs. 99/12765-2 and 05/00587-5) and CNPq (proc. 301303/06-1) and MGC thanks FAPESP (proc. 2009/16799-2) and CNPq for post-doc grant support. MMB acknowledges support from the Medical Research Council (U105185859). CHC thanks FAPESP (proc. 11/22639-8) for financial support.

References

1. Schroeder M. Plos Biology. 2004; 2
2. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, Zhang H, T. F. Consortium. Nucleic Acids Research. 2009; 37:D555–D559. [PubMed: 18948289]
3. Lecuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, Hughes TR, Tomancak P, Krause HM. Cell. 2007; 131:174–187. [PubMed: 17923096]
4. Adryan B, Teichman SA. Bioinformatics. 2006; 22:1532–1533. [PubMed: 16613907]
5. Ernst J, Bar-Joseph Z. BMC Bioinformatics. 2006; 7
6. Choi JK, Yu U, Yoo OJ, Kim S. Bioinformatics. 2005; 21:4348–4355. [PubMed: 16234317]

7. Stuart JM, Segal E, Koller D, Kim SK. Science. 2003; 302:249–255. [PubMed: 12934013]
8. Campiteli MG, Soriani FM, Malavazi I, Kinouchi O, Pereira CA, Goldman GH. BMC Bioinformatics. 2009; 10
9. Jin L, Lloyd R. Journal of clinical laboratory analysis. 1997; 11:2–9. [PubMed: 9021518]
10. Grasso L, Negri A, Foret S, Saint R, Hayward D, Miller D, Ball E. Developmental Biology. 2011; 353:411–419. [PubMed: 21338599]
11. Zhu X, Mahairas G, Illies M, Cameron RA, Davidson EH, Etensohn C. Development. 2001; 128:2615–2627. [PubMed: 11493577]
12. Tabara H, Motohashi T, Kohara Y. Nucleic Acids Research. 1996; 24:2119–2124. [PubMed: 8668544]
13. Puniyani K, Faloutsos C, Xing EP. Bioinformatics. 2010; 26:147–156. [PubMed: 19837720]
14. Mace DL, Varnado N, Zhang W, Frise E, Ohler U. Bioinformatics. 2010; 26:761–769. [PubMed: 19942587]
15. Frise E, Hammonds AS, Celniker SE. Molecular Systems Biology. 2010; 6
16. Barabasi A-L, Oltvai ZN. Nature Reviews Genetics. 2004; 5:101–113.
17. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Nature Genetics. 2005; 37:382–390. [PubMed: 15778709]
18. Carter SL, Brechbhlér CM, Griffin M, Bond AT. Bioinformatics. 2004; 20:2242–2250. [PubMed: 15130938]
19. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. Molecular systems biology. 2007; 3
20. Stuart JM, Segal E, Koller D, Kim SK. Science. 2003; 302:249–255. [PubMed: 12934013]
21. Gilbert, S. Developmental Biology. 6th edn. Sinauer Associates; Sunderland (MA): 2000.
22. da F Costa, L.; Cesar, RM, Jr. Shape analysis and classification. CRC Press; 2001.
23. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. BMC Bioinformatics. 2006; 7
24. Tram, U.; Riggs, B.; Sullivan, W. Encyclopedia of Life Sciences. Macmillan Reference Ltd; London: 2002. Cleavage and Gastrulation in *Drosophila* embryos.

Fig. 1.

Pipeline illustrating the method proposed. Images are filtered and standardized according to orientation and size. These images are then processed in order to obtain feature vectors that represent the concatenation of the patterns of expression on the AP and DV axes. Vectors containing the patterns of spatial expression are compared to construct the co-expression networks. These steps are detailed in the respective following sections.

Fig. 2.

Projection of feature vectors into first and second principal variables for stages 2 to 5. The percentage of explained variance for stage 2 (Figure 2(a)) is 44%, stage 3 (Figure 2(b)) is 40%, stage 4 (Figure 2(c)) is 46% and stage 5 (Figure 2(d)) is 50%.

Fig. 3. Networks obtained for the developmental stages 2 (Figure 3(a)), 3 (Figure 3(b)), 4 (Figure 3(c)) and 5 (Figure 3(d)). The colored subnetworks are enriched for the corresponding annotation.

Table 1

Number of genes represented in each developmental stage

	stg 1	stg 2	stg 3	stg 4	stg 5
Nr. of genes	6	108	113	131	107