Automating Gene Expression Annotation for Mouse Embryo

Liangxiu Han[1], Jano van Hemert[1], Richard Baldock[2], and Malcolm Atkinson[1]

[1] National eScience Centre, School of Informatics, University of Edinburgh, UK
{liangxiu.han,j.vanhemert}@ed.ac.uk, mpa@nesc.ac.uk
[2] MRC Human Genetics Unit, Institute of Genetic and Molecular Medicine, Edinburgh, UK
Richard.Baldock@hgu.mrc.ac.uk

Abstract. *It is of high biomedical interest to identify gene interactions and networks that are associated with developmental and physiological functions in the mouse embryo. There are now large datasets with both spatial and ontological annotation of the spatio-temporal patterns of gene-expression that provide a powerful resource to discover potential mechanisms of embryo organisation. Ontological annotation of gene expression consists of labelling images with terms from the anatomy ontology for mouse development. Current annotation is made manually by domain experts. It is both time consuming and costly. In this paper, we present a new data mining framework to automatically annotate gene expression patterns in images with anatomic terms. This framework integrates the images stored in file systems with ontology terms stored in databases, and combines pattern recognition with image processing techniques to identify the anatomical components that exhibit gene expression patterns in images. The experimental result shows the framework works well.*

Key words: *Gene Expression, Mouse Embryo, Pattern Recognition, and Wavelet Transform*

1 Introduction

Understanding the role of the expression of a given gene and interactions between genes in a mouse embryo requires monitoring the gene expression levels and spatial distributions on a large scale. The availability of high throughput instruments such as RNA in situ hybridization (ISH) method provides the possibility to construct a transcriptome-wide atlas of mouse embryos that can provide spatial gene pattern information for comprehensive analysis of the gene interactions and developmental mechanisms of the mouse embryo. The ISH employs probes to detect and visualise spatio-temporal gene patterns in tissues. The outputs of the ISH on tissues are images stained to reveal the presence of gene expression patterns. To understand gene functions and interactions of genes in depth, we need to transform the raw image data into knowledge. Annotating the raw images of the ISH provides a powerful way to address this issue. The

process of annotating gene expression pattern is to label images with terms from the ontology for mouse anatomy development. If an image is tagged with a term, it means that the anatomical component is expressing as a gene.

Much effort has been invested into the curation of gene expression patterns in developmental biology, for example, the EUREXPress-II project [1] has built a transcriptome-wide atlas database for the developing mouse embryo established by ISH, which has collected more than 18,000 genes at one development stage of the mouse embryo and curated 4 Terabytes of images. The research work in [2] has produced 3375 genes for Genome-wide analysis on Drosophila. Many other gene expression pattern images generated via ISH such as flybase [3] and mouse atlas[4] also provide rich information for the genetic analysis on tissues. The current annotations of gene expressions are made manually by domain experts. With massive amount of curated images available for analysis, it is a huge task for domain experts. Therefore, developing efficiently automatic annotation technique is important. Some existing work [5][6][7] [8] has made attempts on the automating annotation of the gene expression patterns on fruit fly and mouse brain [9]and has provided potential opportunities for further genetic analysis. However, to date, no related work has been done on the automatic annotation of gene expressions for mouse embryos. Comparing with a fly embryo, a mouse embryonic structure [15][16] is more complicated and has more anatomic components, for example, the EURExpress data have 1,500 anatomical features used for the annotations of the mouse embryo.

In this paper, we have used image data from the EURExpress-II project [1] and proposed a new data mining framework for automatic annotation of gene expression patterns in images from developmental mouse embryos. The initial result from the pilot is promising and encouraging. The main contribution of our work consists of following aspects: (1) The combination of statistical pattern recognition and image processing methods can reduce the cost for processing large amount of data and improve the efficiency. We employ the image processing method to standardise and denoise images. The wavelet transform is used to generate and project features from spatial domain to wavelet domain. Considering the high-dimensional features, we use Fisher Ratio analysis to extract the significant features and build up the classifiers based on Linear Discriminant Analysis(LDA). Our classifiers have been evaluated with multi-objective gene expression patterns coexisting in images and the initial results have shown our proposed framework functioned well. (2) Due to multi-anatomical components coexisting in images, this is a typical multi-class classification problem. In this framework, we have formulated this multi-class classification into a two-class problem. We have trained one classifier for each anatomical component. As a result, multi-classifiers for multi-components have been constructed. Each classifier in our framework is a binary classifier, which will give an answer either 'yes' or 'no' when an un-annotated image is coming through. The main advantage is a strong extensibility of the framework. If a new anatomical component to be annotated appears, we can create a new classifier and directly plug it in and no need to train previous existed classifiers. The classification performance

will not affected due to introducing a new class under the same observation dataset. Meanwhile, this design can also improve the scalability and parallel process capability. Classifiers can be arbitarily assembled and deployed based on requirements.

The rest of this paper is organised as follows: the problem domain analysis is described in Section 2; Section 3 presents the methodology used in this proposed framework; Section 4 describes the evaluation result; Section 5 presents the conclusion and future work.

2 Problem Domain Analysis

Currently, in the EURExpress database, 80% of images (4 Terabytes in total) have been manually annotated by human domain experts. For cost-effectiveness, our goal is to automatically perform annotation by classifying the remaining 20% into the correct terms of anatomical components (this would be still 85,824 images to be annotated with a vocabulary of 1,500 anatomical terms). In addition if this is successful we can also validate existing annotations to find errors and inconsistencies. This is a significant challenge.

- Firstly the images generated via ISH include variations arising from natural variation in the source embryos and experimental processing variation and distortion. The same anatomic components therefore may have variable shape, location and orientation.
- Secondly, each image for a given gene will in general be annotated with multiple anatomic terms. This means features for multiple anatomy components coexist in the image, which increases the difficulty of discrimination.
- Thirdly, the number of images associated with a given anatomy terms is uneven. Some of terms may be associated with many images others with only a small number.
- Finally, the dimensionality of each image is high and represented as pixels $m * n$. and in the EurExpress case typically 3Kx4K pixels.

To address these challenges, we propose a new extensible data mining framework that integrates both the images in the file systems and annotation databases and combines image processing with statistical pattern recognition techniques to automatically identify gene expressions in images, as shown in Fig. 1.

To automatically annotate the remained 20% images, we need to learn these annotations by machines first and then automate the classification process by the deployment of classifiers. This would require a training stage to train these annotated data and build up classifiers, a test and evaluation stage for evaluating the performance of classifiers and then finally a deployment stage for deploying the classifiers to perform the classification of un-annotated images.

The processes in the training stage include image integration, image processing, feature generation, feature selection and extraction, and classifier design.

- **Image integration:** Before starting the data mining, we need to integrate data from different sources: the manual annotations have been stored in the

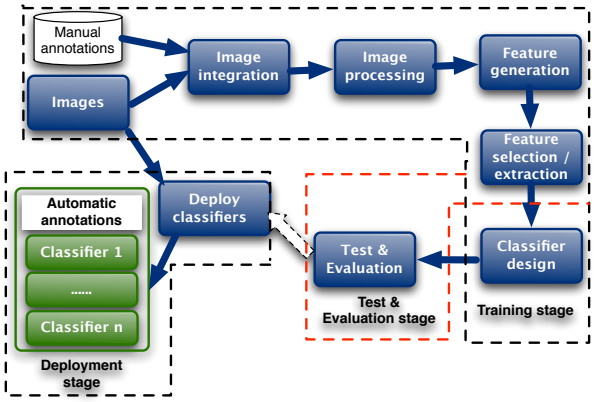


Fig. 1. The data mining framework of automating annotation of gene expressions

database and the images are located in the file system. The outputs of this process are images with annotations.

- **Image processing:** The size of the images is variable. We apply median filtering and image rescaling to reduce image noise and rescale the images to a standard size. The outputs of this process are standardised and denoised images, which can be represented as two-dimensional arrays $(m * n)$.
- **Feature generation:** After image pre-processing, we generate those features that represent different gene expression patterns in images. We use wavelet transform to obtain features. The resulting features of wavelet transform are 2 dimensional arrays $(m * n)$.
- **Feature selection and extraction:** Due to the large number of features, the features need to be reduced and selected for building a classifier. This can be done by either feature selection or feature extraction or both. Feature selection selects a subset of the most significant features for constructing classifiers. Feature extraction performs the transformation on the original features for the dimensionality reduction to obtain a representative feature vectors for building up classifiers.
- **Classifier design:** The main task in this case is to classify images into the right gene terminologies. The classifier needs to take an image's features as an input and for each of anatomical features outputs a rating as 'not detected', 'possible', 'weak', 'moderate' or 'strong' (In the current experimental stage, we use two types ' detected as a gene' and 'not detected as a gene'). We have built separate classifiers for each of anatomical components and considered them independently.

The test and evaluation stage will use the result from the training stage to test images. During this stage, k-fold cross validation is used for evaluating the classification performance. With k-fold validation, the sample dataset is randomly split into k disjoint subsets. For each subset, we train a classifier using the data in the other k-1 subsets and then evaluate the classifier's performance on the data in that subset. Thus, each record of the data set is used once to eval-

uate the performance of a classifier. If 10-fold validation is used, we can build 10 classifiers each trained on 90% of the data and each evaluated on a different 10% of the data.

The deployment stage will deal with the configuration on how to deploy classifiers onto the system, apply classifiers to automatically perform annotation on un-annotated images, and deliver results to the users.

In the following sections, we will mainly focus on the major methods used in the training stage and evaluation stage because of their importance.

3 The methodology

3.1 Feature generation using wavelet transform

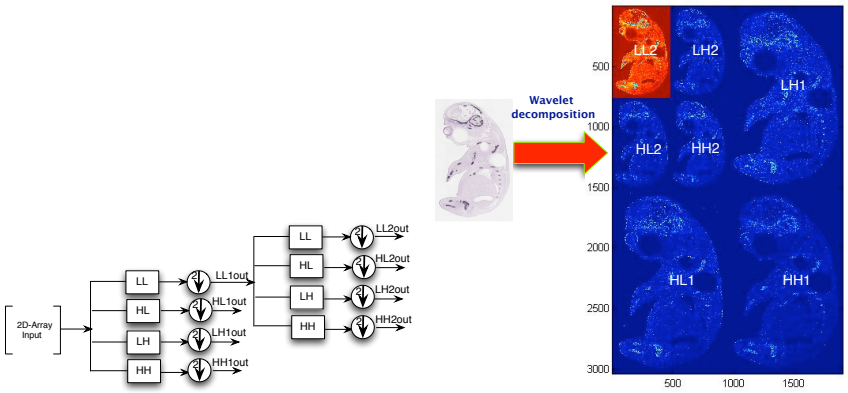
We first obtain samples by integrating both images and manual annotations using a database SQL query to specify which images should be processed. These sample images are filtered and standardised in a uniform size suitable for the feature generation process.

To characterise multi-gene expression patterns in an embryo image, in this paper, we use wavelet transform to represent and generate features. Wavelet transform has been well-recognised as a powerful tool for applications in signal and image processing [10][11][12]. There are two major reasons for using the wavelet transform in our case: (1) Wavelet transform provides a mathematical tool for the hierarchical decomposition of functions, which can decompose images into space and frequency domains, obtain a projective decomposition of the data into different scales and therefore provide local information of images, unlike Fourier transform that only provides global information of images in frequency domain. (2) By using wavelet transform, the image can be decomposed into different subimages at subbands (different resolution levels). The resolutions of the subimages are reduced. On the other hand, the computational complexity will be reduced by operating on a lower resolution image.

In mathematics, wavelet transform refers to the representation of a signal in terms of a finite length or fast decaying oscillating waveform (known as the mother wavelet). This waveform is scaled and translated to match the input signal. In formal terms, this representation is a wavelet series, which is the coordinate representation of a square integrable function with respect to a complete, orthonormal set of basis functions for the Hilbert space of square integrable functions. The wavelet transform includes continuous wavelet transform and discrete wavelet transform. In this case, 2D discrete wavelet transform has been used to generate features from images.

In fact, wavelet transform of a signal can be represented as an input passing through a series filters with down sampling and deriving output signals based on scales (resolution levels). This can be done by iteration process. Fig. 2(a) shows the filter representation of wavelet transform on a 2D array input. LL is a low-low pass filter that is a coarser transform of the original 2D input and a circle with an arrow means down sampling by 2; HL is a high-low pass filter that

transforms the input along the vertical direction; LH is a low-high pass filter that transforms the input along the horizontal direction ; and HH is a high-high pass filter that transforms the input along the diagonal direction. At the first iteration of applying these filters into the input (called wavelet decomposition), the result of wavelet transform will be $LL1out$, $HL1out$, $LH1out$, $HH1out$. At the second iteration, we can continue performing wavelet transformation on $LL1out$ and the output will be $LL2out$, $HL2out$,$HL2out$, $HH2out$. These steps can be continuously and the initial input signal therefore is decomposed into different subbands.



(a) Wavelet decomposition on 2D-array (b) Wavelet decomposition on an image

Fig. 2. Wavelet decomposition

Mathematically, for a signal $f(x, y)$ with 2D array$(M * N)$, the wavelet transform results of applying filters at different resolution levels (e.g., $LL1out$, $HL1out$, $LH1out$, $HH1out$, $LL2out$, $HL2out$, ...) can be calculated as follows:

$$W_\phi(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y)\phi_{j_0,m,n}(x, y). \tag{1}$$

$$W^i_\psi(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y)\psi^i_{j,m,n}(x, y), i = (H, V, D). \tag{2}$$

where $W_\phi(j_0, m, n)$ is $LLout1$ and $W^i_\psi(j, m, n)$ respectively represents $HL1out$, $LH1out$ and $HH1out$ when the wavelet decomposition is performed along the vertical, horizontal and diagonal direction. j_0 is a scale as start point. $\phi_(j_0, m, n)$ and $\psi_{j,m,n}$ are wavelet basis functions. In this case, we use Daubecheis wavelet basis functions(db3) [13].

An example of wavelet transform on an embryo image at the second resolution level is shown in Fig. 2(b). The image is decomposed into four subbands (sub-images). The subbands $LH1$, $HL1$ and $HH1$ are the changes of the image

along horizontal, vertical directions and diagonal directions with the higher frequency component of the image, respectively. After applying filters, the wavelet transform of $LL1$ is further carried out for the second level resolution as $LL2$, $LH2$, $HL2$ and $HH2$. If the resolution of the image is 3040x1900, the sizes of subimages downsampling by 2 at the second resolution level are respectively $LL2$(760x475) , $LH2$(760x475), $HL2$(760x475), HH(760x475), $LH1$(1520x950), $HL1$(1520x950) and $HH1$(1520x950). The total wavelet transform coefficients (features) for the image are 3040*1900=5,776,000.

3.2 Feature selection and extraction using Fisher Ratio Analysis

Due to the resulting of high-dimentional features generated it is necessary to select the most discriminating features. We use Fisher ratio analysis [14] for feature selection and extraction. The Fisher ratio finds a separation space for discriminating features of two classes by maximizing the difference between classes and minimising within the class.

Assuming two classes, $C_1\{x_1, ..., x_i, ...x_n\}$ and $C_2\{y_1, ..., y_i, ...y_n\}$, the Fisher ratio is defined as the ratio of class-to-class variance to the variance of within classes. The Fisher Ratio can be represented as follows:

$$FisherRatio = \frac{(m_{1,i} - m_{2,i})^2}{(v_{1,i}^2 + v_{2,i}^2)}. \tag{3}$$

where $m_{1,i}$ represents the mean of samples at the i^{th} feature in C_1, $m_{2,i}$ represents the mean of samples at the i^{th} feature in C_2. $v_{1,i}$ represents the variance of samples at the i^{th} feature in C_1. Similarly, $v_{2,i}$ represents the variance of samples at the i^{th} feature in C_2.

3.3 Classifier building using LDA

We train each classifier for each anatomical component, and formulate our multiclass problem as a two-class problem. Namely, we treat and divide our sample dataset into two classes during each training: one class contains all of samples with a certain gene expression to be annotated and the other contains all of samples without that gene expression. In this case, we use Linear Discriminant Analysis(LDA) [14] for solving our classification problem. For a given two-class problem ($C_1\{x_1, ..., x_i, ...x_n\}$ and $C_2\{y_1, ..., y_i, ...y_n\}$), the linear discriminant function can be formulated as follows:

$$f(X) = W^t X + w_0. \tag{4}$$

The goal is to find W (weight vector) and w_0 (threshold) so that if $f(X) > 0$, then X is C_1 and if $f(X) < 0$ then X is C_2. The idea is to find a hyperplane that can separate these two classes. To achieve the goal, we need to maximise the target function denoted as follows:

$$T(W) = \frac{|W^t S_B W|}{|W^t S_W W|}. \tag{5}$$

where S_W is called the within-class scatter matrix and S_B is the between-class scatter matrix. They are defined respectively as follows:

$$S_B = (m_1 - m_2)(m_1 - m_2)^t. \tag{6}$$

where,
m_1 = mean of $x_i \in C_1$ and m_2 = mean of $y_i \in C_2$.

$$S_W = S_1 + S_2. \tag{7}$$

where,
$S_1 = \sum_{x \in C_1}(X - m_1)(X - m_1)^t$ and $S_2 = \sum_{y \in C_2}(Y - m_2)(Y - m_2)^t$.

4 Evaluation

We have implemented and deployed our data mining framework into our testbed (a distributed environment). Two databases were created: one for annotations of anatomical components and the other one for feature parameters that is used to store parameters and results from the processes of feature generation and extraction and classifier building. All of image files are located in a file system. Because the features generated are big, we store the features into files hosted in a file system, with references in annotation and parameter databases. Considering the large-scale data mining application in this case, 4 Terabytes data we have curated, we have modularized functional blocks shown in Fig. 1 in order to parallelize these processes in further experiments in near future.

Currently, we have built up 9 classifiers for 9 gene expressions of anatomical components(Humerus, Handplate, Fibula, Tibia, Femur, Ribs, Petrous part, Scapula and Head mesenchyme) and have evaluated our classifiers with multi-gene expression patterns in 809 images. We use the cross validation with 10 folds. The dataset (809 image samples) is divided into 10 subsets. 9 subsets are formed as a training set and one is viewed as a test set. The classification performance is computed based on the average correct or error rate across all 10 tries. The advantage of this method is every sample will be in a test set only once and 9 times in a training set.

The preliminary result of the 10-fold cross validation in our case is shown in table 1. The result shows the correct rate for identifying images with Humerus can achieve 75.25% and the correct rate for identifying images without Humerus gene expression can achieve 79.21%. Similarly, the correct rates for identifying with and without gene expressions on Handplate as 71.05% and 72.31% ; on Fibula as 72.73% and 71.8%; on Tibia as 74.67% and 71.8%; on Femur as 72.41% and 73.45%; on Ribs as 56.14% and 75.38%; on Petrous part as 79.03% and 75.38%; on Scapula as 78.82% and 55.07%. Except the ribs, all other gene expression can be identified well. The various morphologies and the number of ribs in images cause the lower identification rate.

Table 1. The preliminary result of classification performance using 10-fold validation

Classification Performance / Gene expression	Sensitivity	Specificity
Humerus	0.7525	0.7921
Handplate	0.7105	0.7231
Fibula	0.7273	0.718
Tibia	0.7467	0.7451
Femur	0.7241	0.7345
Ribs	0.5614	0.7538
Petrous part	0.7903	0.7538
Scapula	0.7882	0.7099
Head mesenchyme	0.7857	0.5507
Note: Sensitivity: true positive rate. Specificity: true negative rate.		

5 Conclusion and Future Work

In this paper, we have developed a new data mining framework to facilitate the automatic annotation of gene expression patterns of mouse embryos. There are several important features of our framework: (1) the combination of statistical pattern recognition with image processing techniques can help to reduce the cost for processing large amount of data and improve the efficiency. We have adopted the image processing method to standardise and denoise images. Wavelet transform and Fisher Ratio techniques have been chosen for feature generation and feature extraction. The classifiers are constructed using LDA. (2) For enhancing the extensibility of our framework, we formulate our multi-class problem into a two-class problem and design our classifiers with a binary status:'yes' or 'no'. One classifier only identifies one anatomical component. Classifiers for each gene expression are independent on each other. If new anatomical component need be annotated, we do not have to train previous classifiers again. The classifiers can be assembled and deployed into the system based on user requirements. (3) We have evaluated our proposed framework by using images with multi-gene expression patterns and the preliminary result shows our framework works well for the automatic annotation of gene expression patterns of mouse embryos.

The future work will focus on the improvement of the classification performance and parallelise each functional block proposed in this framework in order to enhance the scalability for processing large-scale data of this case in further experiments later on.

Acknowledgments. This work is supported by the ADMIRE project, which is funded by EU Framework Programme 7 FP7-ICT-215024.

The authors acknowledge the support Lalir Kumar and Mei Sze Lam of the EurExpress team (FP6 funding) at the MRC Human Genetics Unit. The authors would also like to thank Jianguo Rao, Jeff Christiansen and Duncan Davidson at the MRC Human Genetics Unit, IGMM, for useful discussions and suggestions.

References

1. EURexpress II project, Retrieved on 08,March, 2009. http://www.eurexpress.org/ee_new/project/index.html
2. Lecuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T. R., Tomancak, P., and Krause, H. M.: Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. Cell 131, 131, 174–187 (2007)
3. Drysdale, R.: FlyBase : a database for the Drosophila research community. Methods Molec. Biol. 420, 45–59(2008)
4. Lein, E. S., Hawrylycz, M. J. et al.: Genome-wide atlas of gene expression in the adult mouse brain. Nature, 445, 168–176(2006)
5. Grumbling, G., Strelets, V., and Consortium, T. F.: FlyBase: anatomical data, images and queries. Nucleic Acids Research. 34, D485-D488(2006)
6. Harmon, C. L., Ahammad, P., Hammonds, A., Weiszmann, R., Celniker, S. E., Sastry, S. S., and Rubin, G. M.: Comparative Analysis of Spatial Patterns of Gene Expression in Drosophila melanogaster Imaginal Discs. LNCS, vol 4453, pp.553-547.Springer, Heidelberg (2007)
7. Pan, J.Y., Balan, A. G. R., Xing, E. P, Traina, A. J. M. and Faloutsos C.: Automatic Mining of Fruit Fly Embryo Images. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 693–698. ACM, NewYork(2006)
8. Zhou, J., and Peng H.: Automatic recognition and annotation of gene expression patterns of y embryos. Bioinformatics 23 (5), 589–596(2007)
9. Carson, J. P., Ju, T., Lu, H.-C., Thaller, C., Pallas, S. L., Crair, M. C., Warren, J., Chiu, W., and Eichele, G.: A Digital Atlas to Characterize the Mouse Brain Transcriptome. PLoS Comput Biology, 1, 0290–0296(2005)
10. Jawerth, B., and Sweldens, W.: An Overview of Wavelet Based Multiresolution Analyses. SIAM Review 36(2), 377-412
11. Stollnitz, E., DeRose, T., and Salesin, D.: Wavelets for Computer Graphics. Morgan Kaufmann Publishers, Inc.(1996)
12. Mallat, S. G.: A Wavelet Tour of Signal Processing.Academic Press(1999)
13. Daubechies, I.: Ten Lectures on Wavelets. S.I.A.M. (1992)
14. Duda, R. O., and Hart, P. E.: Pattern Classification and Scene Analysis. Wiley(1973)
15. Baldock, R., Bard, J., Burger, A. et al.: EMAP and EMAGE: A Framework for Understanding Spatially Organised Data. Neuroinformatics 1(4),309–325(2003)
16. Christiansen, J. H., Yang, Y., Venkataraman, S. et al.:EMAGE: A Spatial Database of Gene Expression Patterns during Mouse Embryo Development.Nucleic Acids Research 34, D637(2006)