Automatic Annotation of Drosophila Developmental Stages Using Association Classification and Information Integration

Tao Meng and Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124, USA
E-mail:t.meng@umiami.edu, shyu@miami.edu

Abstract

In current developmental research, one of the challenging tasks is to understand the spatio-temporal gene expression patterns and the relationships among different genes. In situ hybridization (ISH) assay which shows mRNA spatio-temporal expression patterns in cells and tissues directly is currently widely utilized in the bench work. With the increasing of available ISH images, automatic annotation systems are highly demanded. In this paper, an automatic classification system is proposed for annotating the in situ hybridization images with respect to the developmental stages. The embryo is first segmented from the original image, registered and normalized. The segmented embryo image is then divided into 100 blocks from which the pixel intensity and texture features are extracted and discretized. The multiple correspondence analysis (MCA) based association classification approach is proposed to generate classification rules for different stages based on the training data set. The testing instance is classified by applying the rules generated in the training process and a classification coordination module is incorporated to resolve the conflicts utilizing the weights derived from angle values in the MCA procedure. Experimental results show that our proposed method achieves promising results and outperforms other state-of-the-art algorithms.

Keywords: Drosophila Developmental Stage, Association Classification, MCA-based Classification Model

1. Introduction

In the current post genomic era, biomedical researchers are not only interested in the primary sequences of genes but also the functions of individual genes, interactions among different genes, and how these interactions affect gene expression and phenotypes correspondingly. The research of the development of the model organism such as *Drosophila* has shed light on these issues [1]. By using the state-of-the-art techniques, such as DNA microarray [2] and *in situ* hybridization (ISH) techniques [3], the expression patterns of different genes could be captured during developmental stages for a specific species. Currently, there are several ongoing projects which collect the ISH images at a whole-genome scale. For example, the Berkeley Drosophila Genome Project (BDGP) [4] contains around 97000 digital images of the spatio-temporal expression patterns across six developmental stages for over 7000 genes using ISH technologies. Therefore, researchers could track the changes of patterns in different developmental stages.

Within developmental research, expression pattern comparison is the most biologically meaningful when the images from a similar developmental stage range are compared. However, little work has been done for annotating the developmental stages of the embryos. In [5], the authors extracted Gabor features [6] from the sub-block and used the Regularized Uncorrelated Linear Discriminant Analysis (RULDA) to sort 2705 images from the BDGP database into three developmental ranges (1-3, 4-6, 7-8). In [7], they further improved the regularization to develop the LdaPath algorithm for solving the same classification problem. They claimed that the highest accuracy reached 87.19% in their framework. However, the computational cost is very high and the classification is only based on three of the six developmental stage ranges, which limits the usage of their framework. In [8], using the ISH images from the same database, the researchers proposed the framework to first segment four blocks from the original image based on human inspection and extracted Gabor features to represent the texture information of these blocks. After a PCA-based dimension reduction, the multi-class SVM was utilized for classification and the maximum accuracy was 93.27%. Their framework suffers from two

problems. First, it relies on human inspection to select the sub-blocks for further processing. Therefore, it leads to another problem that the blocks are suitable for their specific task, which is to classify the images into four categories, stages 3, 4, 5, and 6. In addition, the previous two frameworks only took into consideration of the texture information of the embryo without considering the relationship between the expression patterns and developmental stages. Currently, there is no framework to classify the ISH images automatically based on the six stage ranges which span the whole process of Drosophila early development.

Multiple Correspondence Analysis (MCA) is a descriptive data analytic technique in multivariate statistics to analyze simple two-way and multi-way tables for more than two variables, containing a measure of correspondence between the rows and columns. By capturing the correlation between the feature value pairs and classes, it has been utilized to generate association classification rules used in binary classifiers in our previous work [9][10]. Experimental results showed that it achieved relatively promising performance in video concept detection.

In this paper, a MCA-based classification system is proposed for annotating the ISH images from the BDGP database for all six developmental stage ranges. Compared with the previous work, the main contributions of our proposed framework are threefold. First, the proposed framework is able to classify the embryo images into the 6 developmental stage ranges, which match the labels in the database. Second, the proposed framework builds a model to represent the spatial expression patterns and considers the correspondence between the expression patterns and the developmental stages by utilizing MCA-based association classification. Last but not least, MCA used in our previous work did only solve the binary classification problem, but this work extends it to address the multi-way classification problem by reusing the information (angle values) from MCA. Experimental results show that our proposed framework outperforms other state-of-the-art frameworks and other classifiers in Weka [11].

2. Proposed Framework

The proposed framework is shown in Fig. 1. The framework consists of the *Data Preprocessing Module* (1) and *Classification and Coordination Module* (2).

In the *Data Preprocessing Module*, the embryo areas are first segmented from the raw images. Afterwards, the segmented embryo images are registered to the same orientation and size (1200x460) before the normalization step. Each embryo image is then divided into 100 blocks for feature extraction. Three-fold cross validation is used in this framework so that the data set is split into a training data set (two thirds of the whole data) and a testing data set (one third of the whole data). The training

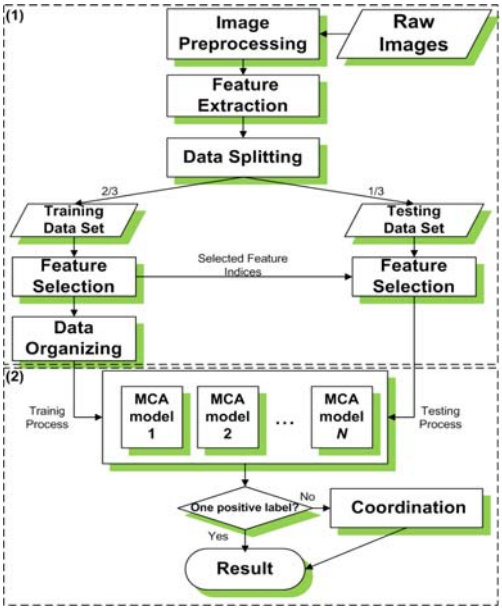

Figure 1. The proposed framework

data set is reorganized to fit the classification module properly. For simplicity of the description, the class label used in the rest of the paper and the corresponding meanings are described here, namely, Class 1 for Developmental Stage Ranges 1-3, Class 2 for Ranges 4-6, Class 3 for Ranges 7-8, Class 4 for Ranges 9-10, Class 5 for Ranges 11-12, and Class 6 for Ranges 13-16.

In the *Classification and Coordination Module*, the MCA model for a certain class is trained and a set of rules of that class is built during the training process. A coordination scheme is developed to make the final decision by reusing the information obtained in MCA in the training process. The classification results are evaluated using accuracy to compare with several existing frameworks.

2.1. Image Preprocessing

The raw images in the BDGP database are of the size (1520x1080) in jpeg format. Because the pictures were taken under the 96-well plate in the experiments, the embryos in the pictures are of different orientations. Due to the fact that the embryo regions have relatively high variance compared with the uniform background, the Otus' method [12] is used to set the threshold for extracting the embryo. After the contour of the embryo images is defined from the previous step, the anterior posterior axis is found by applying principal component analysis (PCA) on the binary images derived from the contour. The embryo images are then rotated so that the anterior posterior axis is aligned horizontally with the

anterior side on the left. Afterwards, a minimum bounding rectangle is calculated for the embryo images and the region within that bounding rectangle is segmented from the grey image generated based on the raw image. Further, all the segmented images are resized to 1200x460 and normalized by applying histogram equalization which uniforms the histogram distribution. Fig. 2 shows an example of the processed image compared with the raw image.

In order to capture the relationship between the expression patterns and the developmental stages, the segmented images are further divided into small blocks and each block carries the information for a certain region. In this study, the images are divided into 100 blocks. The division scheme is shown in Fig. 3 and each block is assigned an ID using a sequential number in a left to right, top to bottom way. Based on our preliminary experimental results, the division scheme represents the local information of each region relatively well.

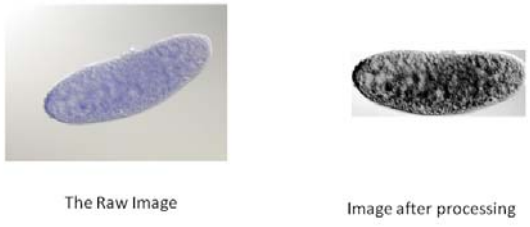

Figure 2. The comparison between the raw image and processed image

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Figure 3. The division scheme

2.2. Feature Extraction

In our proposed framework, the mean pixel value and entropy for each block are extracted. Specifically, they represent the relative expression levels as well as the texture information of a block. In this way, each block is represented by the two-dimensional vector. The vectors of all blocks are concatenated sequentially from Block 1 to Block 100. Therefore, each embryo image is represented by a 200-dimensional feature vector and the spatial information is retained in the sequence of the features.

2.3. Data Splitting, Feature Selection, and Data Organizing

The whole dataset is split to a training data set (2/3 of all data instances) and a testing data set (1/3 of all data instances). The training data set is organized into N training subsets labeled as 1, 2, …, N, where N is the total number of classes in an application. N equals to 6 in this application. Assuming that the total number of instances in the training set is F, the pseudo code of generating the SubSet is given as follows.

SUBSET-GENERATION

```
1   for class_k ← class_1 to class_N
2     t=0;
3     SubSet_k={};
4     NegtiveSet_k={};
5       for instance_i ← instance_1 to instance_F
6         if instance_i is of class_k then
7           SubSet_k ← SubSet_k ∪{instance_i};
8           t ← t + 1;
9         else
10          NegtiveSet_k ← NegtiveSet_k ∪{instance_i};
11        next instance_i
12    SelNegtiveSet_k=select t instances from
      NegativeSet_k randomly;
13    SubSet_k ← SubSet_k ∪ NegativeSe t_k;
14  next class_k
```

Since the feature selection and discretization steps are beyond the scope of this study, the Chi-square feature selection approach and MDL method [13] for discretization implemented in Weka [11] are utilized. In the feature selection step, after computing the ranking scores, the features whose scores are greater than or equal to the sum of the mean value and the standard deviation of all ranking scores are retained. In this study, 45 features are retained and they are sorted by their block ID incrementally to maintain the spatial information. The feature values are then discretized into nominal intervals. The testing data set is discretized using the same intervals derived in the training stage.

2.4. MCA-based Classifier Model

MCA is an extension of the standard correspondence analysis to more than two variables [14] and is applicable for nominal features. The procedure of the MCA is shown in the following example. Supposing there are C data instances in the training data set and E features (E =45 in this study) after feature selection and each feature has H_e (e=1…E) intervals generated from the discretization step.

Let J be the summation of H_1, H_2,…, H_E. Therefore, the indicator matrix (denoted as X) has the dimension of $C \times J$. The Burt Matrix B is calculated using Equation (1).

$$B = X^T X \ . \quad (1)$$

If the grand total of the Burt Matrix is A, the probability matrix P is calculated by dividing each element in B by the scalar A. The vector of the column totals of P forms the vector M. Let D be the diagonal matrix with the elements on the diagonal being the corresponding component in M. By using singular value decomposition (SVD) shown in Equation (2), MCA provides the principal components.

$$D^{-\frac{1}{2}}(P - MM^T)(D^T)^{-\frac{1}{2}} = U \sum V^T \ . \quad (2)$$

The training data are projected to the new principal component space by using the first and second principal components. The correspondence between the feature value pair and positive class label can be represented as the angle between the two vectors. The smaller the angle is, the more correlated the feature value pair to the class is. Because each SubSet_k (k=1…N, where N is the total number of classes) contains only the positive data instances and negative data instances for Class k, it is used to build the MCA model k, which corresponds to Class k. MCA is applied to the training data set to measure the correspondence between each 1-feature value pair and the positive class label. Specifically, each 1-feature value pair and the positive class label are projected to the new principal component space and the angle between them is calculated as the measure of the correspondence between them. Similarly, the correspondence between each 1-feature value pair and the negative class label is also calculated. For example, the feature "mean pixel value" ($Feature_{11}$ or *F11*) of Block 6 is selected and discretized into three intervals from the previous step, labeled as *F11_1*, *F11_2* and *F11_3*. Therefore, this feature has three 1-feature value pairs. By applying MCA to the data instances in training SubSet_3 (i.e., for Class 3), the projections of the 1-feature value pairs could be calculated, as illustrated in Fig. 4. The absolute values of the angles between each 1-feature-value pair and the positive class label are 63.0718, 20.5302, and 156.5062 degrees, respectively. In this example, *F11_1* and *F11_2* have relatively higher correlations with the positive class label; while *F11_3* has a higher correlation with the negative class label.

After calculating the correlation between each 1-feature-value pair and the class label and selecting the threshold value properly based on the training data, a set of one-item rules is generated for both positive and negative classes, respectively. An example one-item rule can be {$Feature_{11}$=*F11_1*} => Class 3. The similar analysis could be generalized to G-feature value pairs ($2 \le G \le E$ and E is the total number of features). The specific procedures for the rule generation and pruning process are described in our previous work [14]. In this paper, the maximum number of G is 8. In this way, the MCA-based Classification Model_k which consists of a set of positive rules and negative rules is built for each SubSet_k (for Class k). It should be pointed out the G-item rules carry the information of the spatial patterns of the embryos. For example, one of the positive rules in Model_1 is: {$Feature_{11}$=*F11_1*, $Feature_{54}$=*F54_3*, $Feature_{77}$=*F77_2*} => Class 1. $Feature_{11}$, $Feature_{54}$, and $Feature_{77}$ are the mean pixel value of Block 6, the entropy value of Block 27, and the mean pixel value of Block 39. It represents the spatial patterns of the embryo image.

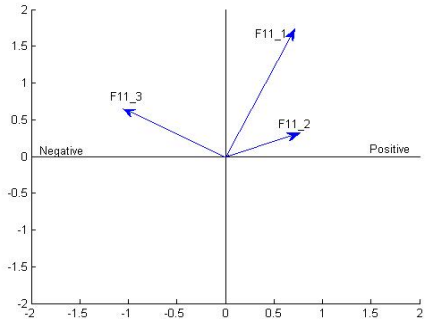

Figure 4. Projections of feature value pair

2.5. Classification and Coordination

After building the MCA-based classification models, the testing procedure is relatively simple. Specifically, let S_k (k=1…N) be the set of positive and negative rules of model_k. The testing instance l is checked to compute the number of positive rules and negative rules it matches for S_k. If the number of matched positive rules is greater than the number of matched negative rules, the testing data instance is labeled positive for Class k and vice verse. If there is a tie, the positive label is assigned to l. Using this method, each testing data instance l is checked per rule set and the corresponding classification results generated forms a sequence R_1, R_2, …, R_N. Ideally, there should be one and only one positive label in the sequence. However, we need to handle two issues under practical conditions: all the values are negative (i.e., no label), or there are positive labels from more than one MCA model (i.e., ambiguity). The coordination module is designed to address these issues.

As described in Section 2.4, each rule carries certain local information of the spatial patterns. Therefore, the global information which is extracted by considering all the nominal features of a data instance together is missing. Given this fact, a new scheme is used in the coordination module. In Section 2.4, the absolute angle value between each 1-feature-value pair and positive class

is calculated for one SubSet. Here, this information is reused to calculate the weight for making the final decision. Assuming in a training SubSet_k (k=1...N) corresponding to Class k, there are E number of features and each feature e has H_e number of different intervals after discretization. If Y_e^a (a=1…H_e) denotes the 1-item feature value pair which *Feature e = Fe_a* and O_e^a is the angle between Y_e^a and the positive class label of Class k, the weight W_e^a is computed using Equation (3) in the training process.

$$W_e^a = 1 - O_e^a / 90 . \quad (3)$$

Because O_e^a is between 0 and 180, the weight value is between 1 and -1. The greater the weight value, the higher the correlation between the 1-item feature value pair and the positive class label is. For a testing data instance l, let each feature value be *f(e)*, which indicates the interval the feature e falls into. For example, if the first feature of l falls into the second interval, then $f(1)$ = 2. The total score is calculated using Equation (4). Here E is the total number of features after feature selection.

$$Score = \sum_{e=1}^{E} W_e^{f(e)} . \quad (4)$$

Now, this *Score* value is used to address the aforementioned issues. If the testing data instance is deemed to be negative by all classification models, the *Score* value will be calculated for each class and the testing data instance is assigned to the label which has the largest *Score*. If the testing data instance is recognized as positive for more than two classes, all these classes are candidate classes and the *Score* values are computed for these candidate classes, respectively. Similarly, the class corresponds to the largest score is assigned to the testing instance. Experimental results show that the proposed coordination module is successful in solving these two issues.

3. Experimental Results

The BDGP database [4] currently contains 97842 images of 7152 genes. Some of the images are out-of-focused and some contain malformed embryos. These images carry little meaningful information and are thus eliminated from the data set. In addition, some images were taken under high resolution and do not contain one intact embryo. These images are also eliminated from the data set because the global information is missing. As the raw images taken represent different views of the embryos and only the comparison of data instances from the same view is meaningful, two sets of data corresponding to the lateral views and dorsal views, respectively, are formed and called "Lateral View Dataset" and "Dorsal View Dataset." After removing those unqualified images and a manually check process, 7471 and 6337 images are selected from the lateral view and dorsal view raw image pool, respectively.

In order to evaluate the performance of our framework, we implemented two other published algorithms for comparison purposes, namely the LdaPath based classification algorithm [7] and support vector machine based algorithm [8]. The parameters used, such as number of scales and the orientations in Gabor filters are tuned to give the best performance on the training data sets. The performance of our proposed framework is also compared with other common classifiers in Weka [11], including C4.5, JRIP, AdaBoost (with C4.5), k-Nearest Neighbors (k=3), Support Vector Machine (with poly kernel). The evaluation criterion is the classification accuracy, which is the percentage of the images in the testing dataset whose classification labels determined by the classifiers match the ground truth. Ten-times three-fold cross validation is used as the testing scheme. The average accuracies and standard deviations for the two data sets are shown in Table 1 and Table 2. In both tables, "MCA-based classifier" indicates our proposed framework, "LdaPath" indicates the framework described in [7], and "Gabor+SVM" is the algorithm implemented based on [8].

From the comparison between the two data sets, it could be seen that the performance of all classifiers on the first data set is better than those of the second one. It matches the fact that the gene expression patterns and texture information are, in general, illustrated more clearly in the lateral view. The LdaPath based classification method which projects the data instances to a new space in order to maximize the inter-class differences and minimize the intra-class differences gives relatively good performance. However, the computational cost is quite high due to the large dimensionality in the feature vector (1280 features are extracted). The method proposed in [8] that extracts texture features from four blocks at the specific positions shows relatively poor performance because it does not take the global expression patterns and texture information into consideration. The method performed well in [8] because it was specifically tuned into their classification task and is not feasible to be generalized to solve other problems The relatively high standard deviation values also indicate this problem. For other classifiers in Weka [11], the AdaBoost+C4.5 algorithm gives relatively good and stable performance. AdaBoost is a meta-algorithm which improves the performance of the classifiers using a multi-step optimization and usually performs well on the data set with few noisy data instances. The images in the database are inspected so that the number of noisy data instances is small, which benefits the AdaBoost algorithm. After all, it shows that our proposed

framework outperforms other frameworks in terms of both the average classification accuracy and the stability of performance. The results indicate that our proposed framework can capture the correspondence between spatial patterns and the developmental stages, and thus is quite useful to improve the classification performance.

Table 1. Performance of lateral view data set

Classification Framework	Average Accuracy	Standard Deviation
MCA-based classifier	**90.14%**	**1.86%**
LdaPath	81.43%	2.14%
Gabor +SVM	77. 86%	5.49%
C4.5	80.07%	2.92%
JRIP	78.73%	2.67%
AdaBoost+C4.5	82.29%	1.96%
3NN	81.26%	4.71%
SVM	82.57%	2.89%

Table 2. Performance of dorsal view data set

Classification Framework	Average Accuracy	Standard Deviation
MCA-based classifier	**87.78%**	**2.27%**
LdaPath	79.76%	3.71%
Gabor +SVM	73.67%	7.53%
C4.5	74.81%	5.46%
JRIP	71.88%	3.12%
AdaBoost+C4.5	81.33%	2.74%
3NN	78.65%	4.82%
SVM	80.88%	3.77%

4. Conclusion

In this paper, a MCA-based multi-class classification framework is proposed to classify the ISH images based on the developmental stages. By using MCA-based correspondence analysis, a set of rules is generated for each class. Each testing data instance is evaluated using the rules. The coordination module is incorporated to address the "no label" and "ambiguity" issues by integrating the information from the previous step. Experimental results show that our proposed framework outperforms several state-of-the-art algorithms and other common classifiers significantly, which demonstrates the effectiveness of our proposed framework.

5. References

[1] P. Tomancak, B.P. Berman, A. Beaton, R. Weiszmann, E. Kwan, V. Hartenstein, et al., "Global analysis of patterns of gene expression during Drosophila embryogenesis," *Genome Biology*, vol. 8, no. 7, R145, July 2007.

[2] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative monitoring of gene expression patterns with complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467-470, October 1995.

[3] L. Jin and R.V. Lioyd, "In situ hybridization: method and applications," *Journal of Clinical Laboratory Analysis*, vol. 11, no. 1, pp. 2-9, January 1997.

[4] P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S.E. Lewis, et al., "Systematic determination of patterns of gene expression during Drosophila embryogenesis," *Genome Biology*, vol. 3, no. 12, December 2002.

[5] J. Ye, J. Chen, Q. Li, and S. Kumar, "Classification of Drosophila embryonic developmental stage range based on gene expresssion pattern images," *Proceedings of the Computational System Bioinformatics (CSB2006).* pp. 293-298, Stanford, CA, USA, August 14-18, 2006.

[6] B.S. Manjunath and W.Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837-842, August 1996.

[7] J. Ye, J. Chen, R. Janardan, and S. Kumar, "Developmental stage annoation of Drosophila gene expression pattern images via an entire solution path for LDA," *ACM Transactions on Knowledge Discovery from DATA (TKDD)*, vol. 2, no 1, pp. 4:1-4:21, April 2008.

[8] H. Zhong, W.-B. Chen, and C. Zhang, "Classifying Fruit Fly early embryonic developmental stage based on embryo in situ hybridization images," *The 2009 IEEE International Conference on Semantic Computing (ICSC2009)*, pp. 145-152, Berkeley, CA, USA, September 14-16, 2009.

[9] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Correlation-based video semantic concept detection using Multiple Correspondence Analysis," *IEEE International Symposium on Multimedia (ISM2008)*, pp. 316-321, Berkeley, CA, USA, December 15-17, 2008.

[10] L. Lin and M.-L. Shyu, "Mining high-level features from video using associations and correlations," *The Third IEEE International Conference on Semantic Computing (ICSC2009)*, pp. 137-144, Berkeley, CA, USA, September 14-16, 2009.

[11] I. H. Witten and E. Frank. *Data Mining Practical Maching Learning Tools and Techniques*, 2nd ed., San Francisco: Morgan Kaufmann, 2005.

[12] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146-165, January 2004.

[13] U. Fayyad and K. Irani , "Multi-interval discretization of continuous-valued attributes for classification learning," *International Joint Conference on Artificial Intelligence - (IJCAI1993*), pp. 1022-1027, France, 1993.

[14] N.J.E. Salkind. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage Publications Inc., 2007.