Mining Spatial Gene Expression Data for Association Rules

Jano van Hemert[1] and Richard Baldock[2]

[1] National e-Science Institute, University of Edinburgh, UK
jvhemert@nesc.ac.uk
[2] MRC Human Genetics Unit, Edinburgh, UK
Richard.Baldock@hgu.mrc.ac.uk

Abstract. We analyse data from the Edinburgh Mouse Atlas Gene-Expression Database (EMAGE) which is a high quality data source for spatio-temporal gene expression patterns. Using a novel process whereby generated patterns are used to probe spatially-mapped gene expression domains, we are able to get unbiased results as opposed to using annotations based predefined anatomy regions. We describe two processes to form association rules based on spatial configurations, one that associates spatial regions, the other associates genes.

Keywords: association rules, gene expression patterns, in situ hybridization, spatio-temporal atlases.

1 Introduction

Association rules are popular in the context of data mining. They are used in a large variation of application domains. In the context of gene expression and images, we can classify previous studies into three categories:

1. Association rules over gene expression from micro-array experiments.
2. Association rules over features present in an image.
3. Association rules over annotated images.

The first category aims to find rules that show associations between genes, and perhaps other things, such as the type of treatment used in the experiment [1,2]. Typical examples of such rules are: if gene a expresses then there is a good chance that gene b also expresses. The second type leads to rules that say something about the relationships between features in images. We found only a few studies in this direction. One which first uses a vocabulary to annotate items found in tiled images and then creates rules that describe the relationships within tiles [3], another which aims to discriminate between different textures by using associations rules [4,5]. Last, the third category extracts rules that show how annotations of images are associated, which is useful to find further images of interest based on an initial set found from a search query [6]. This is the typical concept of "customers who bought this also bought that".

S. Hochreiter and R. Wagner (Eds.): BIRD 2007, LNBI 4414, pp. 66–76, 2007.

We introduce a novel application of mining for association rules in results of *in situ* gene expression studies. In earlier studies in which association rules were applied to gene expression results, these results originated from microarray experiments, where the aim is to find associations between genes [1,2] in the context of broad tissue types. Here in contrast, we will operate on accurate spatial regions with patterns derived from *in situ* experiments. This type of accurate data enables us to extract two types of interesting association rules, first we can extract the same type of relationships between genes. However, we can also extract rules expressed in the form of spatial regions, thereby providing knowledge on how areas in an embryo are linked spatially. The only other study in the direction of spatial association rules the authors are aware of, is solely based on synthetic data [7].

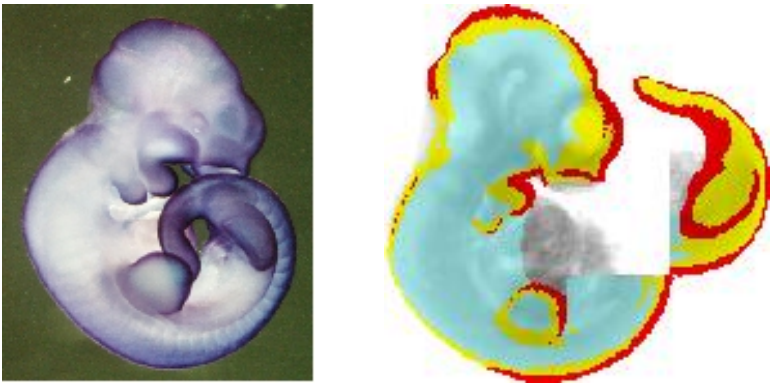
In the next section we describe the Edinburgh Mouse Atlas Project, a spatio-temporal framework for capturing anatomy and gene expression patterns in developing stages of the mouse. Then, in Section 3 we explain the concepts and process of extracting association rules. The spatial framework and association rules are combined in Section 4, which forms the basis for our experiments and results in Section 5. A discussion is given in Section 6.

2 Edinburgh Mouse Atlas Project

EMAGE (`http://genex.hgu.mrc.ac.uk/`) is a freely available, curated database of gene expression patterns generated by *in situ* techniques in the developing mouse embryo [8]. It is unique in that it contains standardized spatial representations of the regions of gene expression for each gene, denoted against a set of virtual reference embryo models. As such, the data can be interrogated in a novel and abstract manner by using space to define a query. Accompanying the spatial representations of gene expression patterns are text descriptions of the sites of expression, which also allows searching of the data by more conventional text-based methods terms.

Data is entered into the database by curators that determine the level of expression in each *in situ* hybridization experiment considered and then map those levels on to a standard embryo model. An example of such a mapping is given in Figure 1. The strength of gene expression patterns are classified either as no expression, weak expression, moderate expression, strong expression, or possible detection.

In this study we restrict to a subset of the data contained in the database. This subset of data originates from one study [9] and contains 1618 images of *in situ* gene expression patterns in a wholemount developing mouse embryo model of Theiler Stages 16, 17, and 18 [10]. The study includes 1030 genes; a subset of genes were screened two or three times. By mapping the strong and moderate expression patterns of these images on to the two-dimensional model for Theiler Stage 17 shown in Figure 1(b), we can work with all these patterns at the same time.



(a) Original image shows expression of Hmgb1 in a mouse embryo at Theiler Stage 17

(b) Standard embryo with mapped levels of expression (red=strong, yellow=moderate, blue=not detected)

Fig. 1. An example of curating data; the gene expression in the original image on the left is mapped on to the standard embryo model of equal developmental stage on the right (entry EMAGE:3052 in the online database).

3 Association Rules

Various algorithms exist to extract association rules, of which the Apriori [11] algorithm is the most commonly used and we too shall use it in this study. It entails a two-step process, defined below, which consists of first generating the set of frequent itemsets, from which association rules are extracted that are above a certain confidence level.

Definition 1

1. *Given a set of items I, the input consists of a set of* transactions *D, where each transaction T is a non-empty subset of items taken from the* itemset *I, so $T \subseteq I$.*
2. *Given an itemset $T \subseteq I$ and a set of transactions D, we define the* support *of T as $support_D(T)$ equals the proportion of transactions that contain T to all transactions $|D|$.*
3. *By setting a* minimum support level *α, where $0 \leq \alpha \leq 1$, we define* frequent itemsets *to be itemsets where $support_D(T) \geq \alpha$.*

Definition 2

1. *An* association rule *is a pair of disjoint itemsets, the antecedent $A \subseteq I$ and the consequent $C \subseteq I$, where $A \Rightarrow C$ and $A \cap C = \emptyset$.*
2. *The concept of support of an association rule carries over from frequent itemsets as $support_D(A \Rightarrow C) = support_D(A \cup C)$.*

3. We define the confidence *of an association rule* $A \Rightarrow C$ *as:*

$$confidence_D(A \Rightarrow C) = \frac{support_D(A \cup C)}{support_D(A)}$$

In other words, frequent itemsets are items that frequently occur together in transactions with respect to some user defined parameter, i.e., the minimum support. In an analog way, the confidence of association rules shows how much we can trust a rule, i.e., a high confidence means that if A occurs, there is a high chance of C occurring with respect to the set of transactions.

The prototypical example to illustrate association rules uses the domain of the supermarket. Here a transaction is someone buying several items at the same time. An itemset would then be something as $\{\text{jam}, \text{butter}, \text{bread}\}$. If this itemset is also a frequent itemset, i.e., it meets the minimum support level, then a possible association rule would be $\{\text{jam}, \text{butter}\} \Rightarrow \{\text{bread}\}$.

We provide the definition of *lift* [12], which is a popular measure of interestingness for association rules. Lift values larger than 1.0 indicate that transactions $(A \Rightarrow C)$ containing the antecedent (A) tend to contain the consequent (C) more often than transactions that do not contain the antecedent (A). Lift is defined as,

Definition 3. $lift(A \Rightarrow C) = confidence(A \Rightarrow C)/support(C)$

4 Applying Association Rules on the Spatial Patterns

We create a set of "probe patterns" by laying a grid over the standard embryo model. Each point in the grid is a square of size 5×5 pixels. The whole image is 268×259 pixels.

To create a relationship between the gene expression patterns in the images, we first build a matrix of similarities between those patterns and the probe patterns. Each element in the matrix is a measure of similarity between the corresponding probe pattern and gene-expression region. We calculate the similarity as a fraction of overlap between the two and the total of both areas. This measurement is intuitive, and commonly referred to as the Jaccard index [13]:

$$\text{similarity}(d_1, d_2) = \frac{S(d_1 \wedge d_2)}{S(d_1 \vee d_2)},$$

where $S(x)$ calculates the size of the pattern x. The pattern $d_1 \wedge d_2$ is the the disjunction or intersection of the image d_1 and the probe pattern d_2, while the pattern $d_1 \vee d_2$ is the union of these regions. The similarity is higher when the overlapping area is large and the non-overlapping areas are small.

In Figure 2, two gene expression pattern images are shown, together with their patterns in the standard embryo model. Also, three examples of probe patterns are shown. We pair probe patterns with gene expression patterns and then calculate the Jaccard Index, which gives us a measure of overlap. A high number means much of the two patterns overlap, where 1.0 would mean the

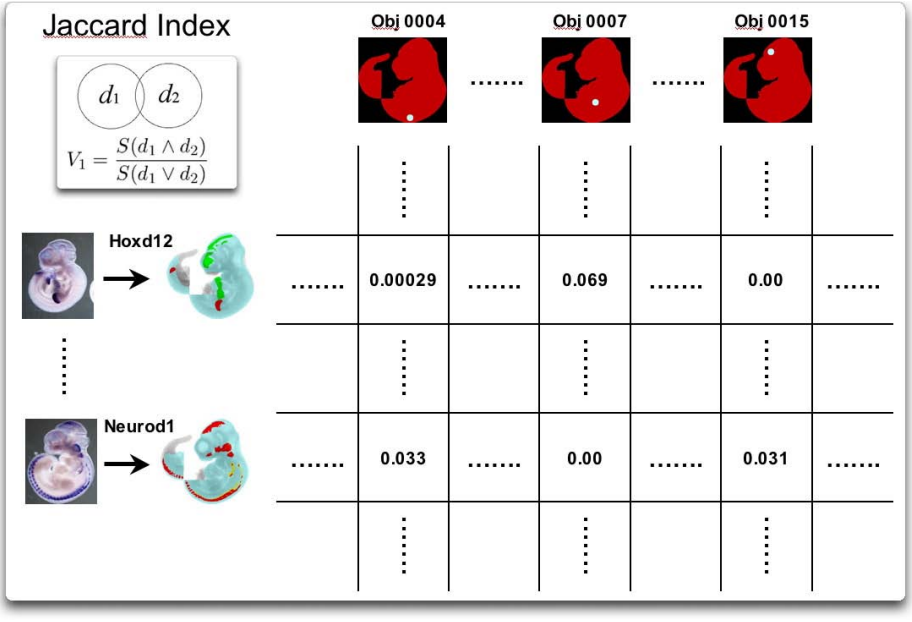


Fig. 2. Similarity matrix. Each original gene-expression assay shown on the left-hand side is mapped on to the standard model embryo, a Theiler Stage 17 wholemount. The probe patterns, shown at the top are then compared with the mapped gene-expression patterns using the Jaccard Index as a similarity measure. These are depicted in the table. The actual simularity matrix has 1618 rows and 1675 columns.

unlikely event of total overlap. A very small number, such as 0.00029, means only a little area of the probe pattern overlaps with a large gene expression pattern, and 0.0 would mean no overlap occurs at all. This is important to note as later we will use a threshold to filter these latter two occurrences.

From the similarity matrix, or *interaction matrix*, two different sets of transactions are constructed, which in turn lead to two different types of association rules.

1. The items I are genes from the data set, where a transaction $T \subseteq I$ consists of genes that all have an expression pattern intersecting with the same probe pattern.
2. The items I are the probe patterns, where a transaction $T \subseteq I$ consists of probe patterns all intersecting with the expression patterns in the same image.

To create the first type of transactions, we take for each probe pattern r, every gene g from which its associated gene expression pattern g_e satisfies the minimum similarity β, i.e., similarity$(r, g_e) > \beta$, to form the itemset.

The second type of transactions is created in a similar way. For each gene expression pattern g in the database we create an itemset that consists of a set

of probe patterns that intersect with the gene expression pattern g_e. Each probe pattern r must satisfy the minimum similarity β, i.e.., similarity$(r, g_e) > \beta$, to get included in the itemset.

When the transactions are formed, the standard Apriori algorithm is used to create frequent itemsets, where the minimum support is set to different values for the different types. The common procedure is to start with a high minimum support, e.g., 90%, which often does not yield any results, and then reduce the threshold as far as the algorithm will support it. At some point either the amount of memory required or the amount of time will render the algorithm useless, at which time we stop and take the results from previous tried minimum support level. From the frequent itemsets we build association rules. Generally we want rules we can be confident about, hence we set the minimum confidence level to 0.97.

5 Experiments and Results

We generated a 1675 square probe patterns of size 5×5 to cover the whole standard embryo model. These parameters were chosen first to match the number of images used in this study. Also, the 5×5 probe patterns allow sufficiently large transactions and a sufficiently number of transactions. When forming transactions, we used a minimum similarity of $\beta = 0.005$. This latter parameter setting was chosen after first performing the whole process with $\beta = 0.00$, which resulted in frequent itemsets and consequently, association rules, with a very high support (above 80%). This generally happens when items are over-represented, and then dominate the analysis. This is caused by gene expression patterns that cover an extremely large area of the embryo. As such patterns will intersect always they do not make an interesting result for finding associations. Also, such association rules are obvious ones.

For example, when using no minimum similarity, e.g., $\beta = 0.00$, and searching for associations by genes, the highest supported association rule is Hmgb2 $\Rightarrow$ Hmgb1 with a support of 0.881, a confidence of 0.993, and a lift of 1.056. These genes are known to be highly associated and are common to many processes [14], hence they express over much of the embryo.

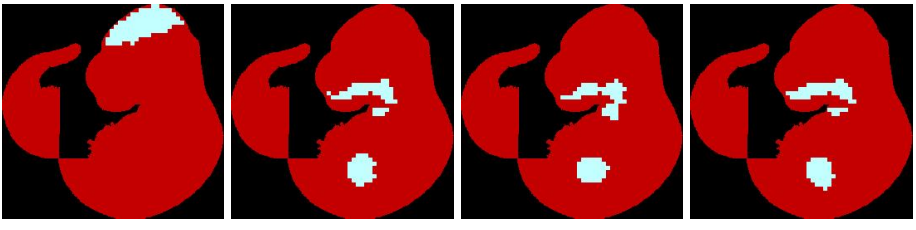
By using the similarity measure to our advantage we can filter out these over-represented genes as in these cases the similarity measure will be small due to the large non-overlapping areas of expression. We found a threshold of $\beta = 0.005$ is sufficient to exclude these patterns. Smaller values rapidly decrease the number of genes left after filtering, and this will make the analysis useless. The runtime of the Apriori algorithm on these data sets is a few seconds on a 2.33Ghz Intel Core Duo chip.

5.1 Associations by Genes

Table 1 shows the association rules found when transactions are genes that have regions of expression intersecting with the same probe pattern. Here a minimum support level of 0.06 is used, and a minimum confidence of 0.97. The lift values

Table 1. Association rules based on itemsets of genes and where a transaction is a set of genes exhibiting expression all intersecting with the same probe pattern, created with a minimum support of 0.06 and a minimum confidence of 0.97.

Rule	*Antecedent*	*Consequent*	*Support*	*Confidence*	*Lift*
1	Lhx4 Otx2	Dmbx1	0.065	0.971	10.871
2	Lhx4 Dmbx1	Otx2	0.065	0.990	9.226
3	Brap Zfp354b	9830124H08Rik	0.060	0.979	10.225
4	9830124H08Rik Trim45	Brap	0.061	0.979	11.813
5	9130211I03Rik Zfp354b	9830124H08Rik	0.062	0.980	10.230



(a) Rule 1 & 2 (b) Rule 3 (c) Rule 4 (d) Rule 5

Fig. 3. Spatial regions corresponding to the conjunction of all transactions, i.e., probe patterns, that lead to each association rule.

of the resulting rules are high; rule 2 for instance has the smallest lift of 9.2. This tells us that for all these rules, if the antecedent occurs, it is far more likely that the consequent occurs than that the consequent does not occur. In other words, these rules are trustworthy under assumption of the provided data.

We can provide more background information about each rule by showing the transactions they were inferred from. As in this case the transactions are the probe patterns, we can display them all at once in the standard embryo model. Figure 3 shows the spatial regions within the embryo model of each rule. Note that rule 1 and rule 2 have the same set of genes and therefore reference the same spatial region. Both these rules associate the genes Lhx4, Otx2, and Dmbx1 in the midbrain. Rules 3, 4, and 5, all associate genes in the forebrain, 1st branchial arch and forelimb. However, rule 4 is quite different to rules 3 and 5, as it has two continuous regions instead of three. Important to note is the accuracy of the expression patterns in the latter three rules. All three rules share common genes, but the differences in genes are cause for subtle changes in the patterns seen in Figure 3.

5.2 Associations by Spatial Regions

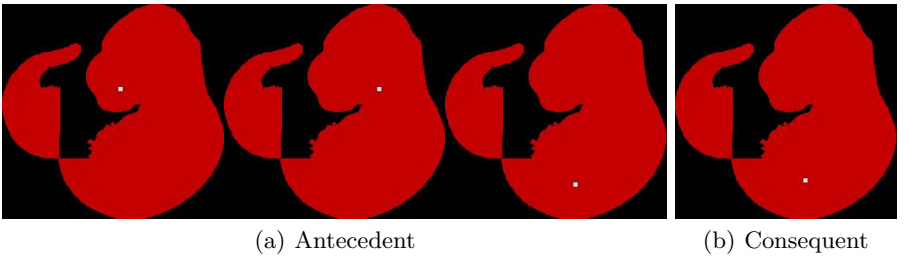
When calculating the association rules based on spatial regions representing items in transactions, which are in turn representing probe patterns intersecting with patterns of the same gene, we end up with millions of rules. The reason for

this is that many rules will state obvious knowledge. This is a common effect in data mining exercises, and it allows one to verify the quality of knowledge extracted. In the context of spatial transactions, most of the rules we extracted state: if a probe pattern exhibits expression of gene g, then it is very likely a neighbouring probe pattern will also exhibit expression of gene g. Although from a data miner's perspective, such rules aid in improving trust in the correctness of the system, they do not provide new information to biologists, as gene expression patterns are inherently local.

The next step, after extracting the association rules, is to select the rules that are newsworthy. As local rules are reasonably obvious, many of these rules state that if a gene exhibits expression in a probe area than it is likely the adjacent probe area also exhibits the same gene. We therefore sort all the rules by the relative distances of the probe patterns in each rule. This way, rules that show relationships between probe patterns further away from each other can be distinguished from association rules operating on local patterns. This sorting uses the sum of the absolute Euclidean distance of every pair of probe patterns in each rule divided by the number of pairs. The higher this sum, the larger the relative distances between pairs of probe patterns. We will show the two rules with the largest average distance between the probe patterns.

Figure 4 shows an association rule based on probe patterns as items. The three square patterns in Figure 4(a) form the antecedent of the rule, while the square pattern in Figure 4(b) forms the consequent of the rule. It has a support of 0.130, which is quite high. The confidence of 1.00 means that if the patterns in the antecedent are expressing a particular gene then under the data provided, the regions in the consequent *must* also exhibit expression of the same gene. The lift is 3.53; this tells us that rules which contain the antecedent will be more likely to also contain the pattern in the consequent than not contain the pattern in the consequent. This confirms the rule provides valuable information about the data.

This rule is extracted from gene expression patterns involving the following list of genes: 9130211I03Rik 9830124H08Rik Abcd1 Ascl3 BC012278 BC038178



(a) Antecedent (b) Consequent

Fig. 4. Spatial region based association rule with a support of 0.130, a confidence of 1.00, and a lift of 3.53. This should be interpreted as "if a gene exhibits expression in all areas shown in (a), then it shall also be exhibiting expressing in (b)". The average Euclidean distance between the probe patterns is 85.0 pixels.

Bhlhb2 Brap Cebpa Chek2 Cited4 Cops6 Creb3l4 Cxxc1 Elf2 F830020C16Rik Fgd5 Gmeb1 Jmjd2a Jrk Mefv Mid2 Mtf1 Nab2 Neurog3 Phtf1 Piwil1 Pole3 Prdm5 Rfx1 Rnf20 Rxra Snf8 Tcf12 Tcfcp2 Thap7 Trim11 Trim34 Wbscr14 Zbtb17 Zfp108 Zfp261 Zfp286 Zfp354b Zfp95.

Figure 5 shows a similar association rule. It has a support of 0.101, a confidence of 1.00, and a lift of 3.53. More of these rules were extracted and show us there is a strong relationship between the areas which project onto the developing eye/forebrain, hindbrain and limb/kidney. It involves the following list of genes: 9130211I03Rik Abcd1 Ascl3 BC012278 BC038178 Bhlhb2 Brap Brca1 Cebpa Chek2 Cited4 Cxxc1 Elf2 F830020C16Rik Fgd5 Gmeb1 Jmjd2a Jrk Mid2 Mtf1 Nab2 Neurog3 Phtf1 Piwil1 Pole3 Prdm5 Rfx1 Rnf20 Snai2 Snf8 Tcf12 Tcfcp2 Thap7 Trim11 Trim34 Trim45 Wbscr14 Zbtb17 Zfp108 Zfp113 Zfp261 Zfp354b Zfp454 Zfp597 Zfp95.

(a) Antecedent (b) Consequent

Fig. 5. Spatial region based association rule with a support of 0.101, a confidence of 1.00, and a lift of 3.53. This should be interpreted as "if a gene exhibits expression in all areas shown in (a), then it shall also be exhibiting expressing in (b)". The average Euclidean distance between the probe patterns is 84.5 pixels.

6 Discussion

We show a novel application of association rules extracted from unique data; accurate spatial localization of gene expression patterns derived from *in situ* experiments. Approaching the data from two different angles allows us to derive two different types of associations. The first type shows relationships between genes, which form a popular field of study for data mining applications. The second type extracts rules that link spatial regions of expression patterns within the embryo. So far, just a few studies consider such rules and they operate on synthetic data. In this study we have shown that this type of analysis can yield potentially interesting and novel associations. The next step will include both automatic (e.g. GO term enrichment and pathway analysis) and manual analysis of the biological meaning of these rules. The identification of spatial associations within the embryo using unbiased sampling is, so far as we can determine, completely novel.

Using our approach of generated probe patterns to represent items in conjunction with a similarity measure between a gene expression pattern and a particular

probe pattern provides two major advantages. It allows to filter out expression patterns that exhibit an area so large they threaten to dominate the frequent itemsets. These expression patterns generally do not add much knowledge as they show ubiquitous gene-expression over almost the whole embryo.

A further advantage is a significant decrease in the influence of the human bias that is present in the annotations of *in situ* images. Similar experiments as reported in this study, but performed on the annotations proved difficult due to a lack of richness, often resulting in support levels too low to consider useful, for instance less than 3%. There are a number of reasons why human annotations lack richness compared to the spatial annotations present in the curated database. Most important is the interest of the annotator as this will result in a focus on particular features present in the image. Then there is the knowledge of the annotator, which will significantly influence what features will be picked up for annotation, and may also result in error. Last we mention time constraints, which always form a major bottleneck in interactive annotations. The result is that manual text annotations are at best low-resolution descriptions of the data but are typically partial and biased observations.

Last, we want to mention the results we showed from tracing back the transactions that are the cause for a certain association rule. In other analyses, such as supermarket data, these transactions are rather meaningless. However, here in the spatial context and in the biological context of genes, these transactions provide supporting evidence to the biologist to understand and verify the consequence of the association rules.

Our near future goal is to provide an interface freely accessible via any web browser to steer the process of the extraction of both types of association rules. By providing a feedback mechanism we then allow biologists to evaluate the usefulness of the rules.

Longer term goals are to include the temporal aspects present in the database, thereby providing rules that state relationships between different stages of development. Also, providing comparative analysis of association rules across species, specifically between the mouse and human model, to highlight similarities and differences in gene and spatial interactions.

References

1. Creighton, C., Hanash, S.: Mining gene expression databases for association rules. Bioinformatics **19**(1) (2003) 79–86
2. Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J., Gandrillon, O.: Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data. Genome Biology **3**(12) (2002)
3. Tešić, J., Newsam, S., Manjunath, B.: Mining image datasets using perceptual association rules. In: Proceedings of SIAM Sixth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Third SIAM International Conference. (2003)
4. Rushing, J., Ranganath, H., Hinke, T., Graves, S.: Using association rules as texture features. IEEE Transactions on Pattern Analyses and Machine Intelligence (2001)

5. Bevk, M., Kononenko, I.: Towards symbolic mining of images with association rules: Preliminary results on textures. Intelligent Data Analysis **10**(4) (2006) 379–393
6. Malik, H., Kender, J.: Clustering web images using association rules, interestingness measures, and hypergraph partitions. In: ICWE '06: Proceedings of the 6th international conference on Web engineering, New York, NY, USA, ACM Press (2006) 48–55
7. Ordonez, C., Omiecinski, E.: Discovering association rules based on image content. In: Proceedings of the IEEE Advances in Digital Libraries Conference (ADL'99), Baltimore, Maryland (1999)
8. Christiansen, J., Yang, Y., Venkataraman, S., Richardson, L., Stevenson, P., Burton, N., Baldock, R., Davidson, D.: Emage: a spatial database of gene expression patterns during mouse embryo development. Nucleic Acids Research **34** (2006) 637–641
9. Gray, P., et al.: Mouse brain organization revealed through direct genome-scale tf expression analysis. Science **306**(5705) (2004) 2255–2257
10. Theiler, K.: The House Mouse Atlas of Embryonic Development. Springer Verlag, New York (1989)
11. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In Buneman, P., Jajodia, S., eds.: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C. (1993) 207–216
12. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In Peckham, J., ed.: Proceedings ACM SIGMOD International Conference on Management of Data. (1997) 255–264
13. Jaccard, P.: The distribution of flora in the alpine zone. The New Phytologist **11**(2) (1912) 37–50
14. Pallier, C., Scaffidi, P., Chopineau-Proust, S., Agresti, A., Nordmann, P., Bianchi, M., Marechal, V.: Association of chromatin proteins high mobility group box (hmgb) 1 and hmgb2 with mitotic chromosomes. Mol. Biol. Cell **14**(8) (2003) 3414–3426