Spatio-temporal framework for integrative analysis of zebrafish development studies

Belmamoune, M.

CHAPTER 6

TOOLS FOR FINDING SPATIO-TEMPORAL PATTERNS OF GENE EXPRESSION DATA IN ZEBRAFISH

Based on:

M. Belmamoune and F. J. Verbeek.
Mining the zebrafish 3D patterns of gene expression database for association rules.
(Submitted, 2009)

Abstract

The analysis and mining of patterns of gene expression provides a crucial approach in discovering knowledge such as finding genetic networks that underpin the embryonic development. In this chapter we describe the extension of the Gene Expression Management System (GEMS) to a framework for data mining and results analysis. As a proof of principle, the GEMS has been equipped with data mining applications suitable for spatio-temporal tracking, thereby generating additional opportunities for data mining and analysis. The analysis of the genetic networks uses spatial, temporal and functional annotations of the patterns of gene expression data stored in GEMS. Combining mining with the available capabilities of GEMS can significantly influence and enhance current data processing and functional analysis strategies.

6.1 Introduction

Data mining techniques are used to identify patterns intrinsic in data, and thereby among other things, support hypothesis generation. It is recognized that the application of data mining techniques involves many tasks supported by a heterogeneous suite of tools. Additionally, interpretation of data mining results requires many decisions taken by experts that must be familiar with data mining techniques and at the same time have sufficient background knowledge on the area under study. These requirements are however, not common to all end-users. Therefore, we propose an embedded framework for both data mining application and results interpretation. In this chapter we present our approach that focuses on embedding mining algorithms on the GEMS framework. The GEMS has been extended to serve as an effective environment of knowledge discovery and interpretation. In the same framework, data mining could be applied and a primary analysis of the discovered rules could also be performed using the patterns annotations, images and links to external resources. We believe that such framework will facilitate data interpretation and analysis.

Gene expression profiles on the level of the transcripts, as well as on the level of the proteins can be a valuable tool to understand gene function. A lot of available methods for gene-expression data-analysis are based on clustering algorithms. These algorithms tend to focus on data with the same expression mode while the transcriptional relation between genes is not addressed. Our attempt to find new patterns in the data was accomplished with association rules. Unlike clustering techniques, this method reveals mutual interaction among genes. In this manner, biologically relevant associations between different genes can be revealed.

In this chapter we discuss our proof of concept methodology that we adopted to facilitate analysis of mining results using association-rule mining technique to discover elements with correlated frequently within our gene expression dataset.

Market Basket Analysis (Agrawal et al, 1993) is a typical and widely-used example of association rule mining. In bio-molecular life sciences research studies, association rules are typically applied to gene expression results obtained from microarray experiments. The first step in microarrays mining procedures is to find association rules between patterns of gene expression. The second step is to find a biological interpretation of the discovered associated patterns. This step is the most delicate and time consuming phase to analyze the discovered rules since the results have to be accurately placed into context with existing biological knowledge, such as scientific literature or sequence data. In our case, we work, on accurate 3D patterns of gene expression that were annotated with standardized and structured metadata during data storage into the GEMS database. The way in which this information is organized makes the interpretation of mining results easier.

6.2 Methods

Association rules discovery is a mining method that has been extensively used in many applications to discover associations among subsets of items from large transaction databases (Agrawal, 1993 et al; Liu, 1998).

Definition:

1. Given a set of items I = {i_1, i_2, i_3, …, i_n} and a set of transactions D = {T_1, T_2, …, T_m}, each transaction T in D is a subset of items in I.
2. Given a set of items (for short *itemset*) X$\subseteq$ I, the support of X is defined by: Support(X) = freq (X)/|D|, which means that the support is equal to the proportion of transactions that contain X to all transactions |D|.
3. An association rule has the following implication form:

a. $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The itemsets X and Y are called *antecedent* (Left-Hand-Side or LHS) and *consequent* (Right-Hand-Side or RHS) of the rule.

4. Each rule is associated with its confidence and support:

Confidence ($X \Rightarrow Y$) = freq ($X \cup Y$)/freq (X), support ($X \Rightarrow Y$) = support ($X \cup Y$) where support ($X \cup Y$) = freq ($X \cup Y$)/$|D|$.

Given a set of transactions (the database), mining for association rules is to discover all association rules that have support and confidence greater than the user specified minimum support and minimum confidence. In general, an association mining algorithm works in two steps. First all itemsets that satisfy the minimum support are generated. Second, generation of association rules that satisfy the minimum confidence using the large itemsets. An itemset is simply a set of items and a large itemset is an itemset that has transaction support higher than the minimum.

The prototype example to illustrate association rules uses the domain of the supermarket (Agrawal et al, 1993). Here a transaction is someone buying several items at the same time. An itemset would then be something as {*cheese*, *beer*} and an association rule is as follow: *cheese* ⇒ *beer* [support = 10%, confidence = 80%]. This rule says that 10% of customers buy *cheese* and *beer* together and those that buy *cheese* also buy *beer* 80% of the time.

There are many efficient algorithms to find association rules, major issue remains to find the right algorithm to meet our needs. We began our gene expression mining studies with the APRIORI algorithm. We took this algorithm since it is the basic algorithm for association-rule mining. APRIORI was extensively studied and successfully applied in many problem domains (Agrawal et al., 1993, 1994). It depends on a very basic property, i.e. for an itemset to be frequent; each of its subset must also be a frequent itemset. The algorithm starts with a single item in the set and then runs iteratively with each frequent itemset detected in the previous level increases by one. This algorithm has many advantages like the capability to find frequent patterns, accuracy and controlled candidate generation. However, it has some limitations. Normally different genes have different temporal expression. Some genes are expressed more frequent and earlier in time then others. Thus considering only the occurrence count of each item (gene) may not lead to a fair measurement. Therefore we moved to the Progressive Partition Miner algorithm (PPM) (Lee et al, 2001) that we apply on our set of data. The idea of PPM algorithm is to first partition a dataset and then progressively accumulates the occurrence count of each itemset based on the intrinsic partitioning characteristics. The PPM algorithm employs a filtering threshold in each partition to early prune those cumulatively infrequent itemsets.

Implementation

We defined and implemented the resources required for the interactive rule mining framework using a platform/language with java as our technology support. (1) We build a java application (cf. Figure 1) that can be executed in two different ways: as an autonomous java agent and through the user interface. Users are able to execute the PPM mining algorithm by sending a HTTP request. (2) The application processes submitted requests and queries the GEMS PostgreSQL database to generate a dataset. The query result is pre-processed to a multi-line text file where each line is considered as a transaction. A transaction is a developmental stage and the items are the expressed genes at this stage. The application runs first to find the frequent 2-genesets in the data. (3) From the frequent 2-genesets the association rules are mined and presented to the user. We provide a graphical user interface to start the mining procedure and to explore the generated rules for data interpretation and analysis.

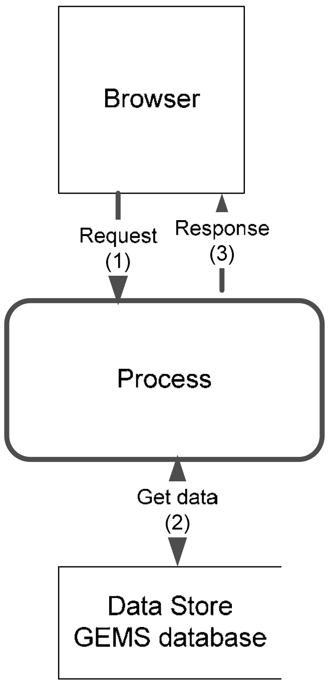

Figure 1: The process flow of the web-application to mine expression patterns.

6.3 Dataset resources

Our case study concerns spatio-temporal patterns of gene expression in zebrafish. Patterns are the result of fluorescent *in situ* hybridization (FISH) experiments and visualized with the Confocal Laser Scanner Microscopy (CLSM). This methodology of patterns generation enables a precise spatial localization of genes expression. This spatial localization enhances extremely functional analysis of genes function. The patterns are subsequently annotated and stored in the GEMS database (cf. chapter 4). We initially analyze the GEMS database using patterns spatio-temporal information. Subsequently, we use the annotations of the patterns supported with the 3D images to post-process the rules that we obtained.

In addition to GEMS data, we used other datasets to first validate and explore the PPM algorithm. We validated the Java application of the PPM algorithm before its integration within the GEMS framework. For this validation, we used the same dataset as presented in (Lee et al, 2001) to get the same mining results. Subsequently, we explored this association-rule technique and we apply it on ZFIN gene expression data (http://zfin.org). We imported ZFIN data in a local database that we query to generate a dataset.

6.4 Results

ZFIN is a large and rich resource of gene expression data. In ZFIN dataset, we found a large amount of rules. To limit the analysis to a small number, we selected these with [support >40, confidence >80] (cf. Table 1). Additionally, for data analysis we limited expression information to these realized under the same experimental conditions (mRNA *in situ* hybridization) and obtained between "prim 15" en "long pec" (stages of development) and (cf. Table 2).

Rule number	**ANTECEDENT**	**CONSEQUENT**
1	*Btg2*	*Tbx20*
2	*Hoxa3a*	*Tbx20*
3	*Hoxa3a*	*Ccnb1*

Table 1: An example extracted from the ZFIN result set using the PPM algorithm (support > 40% and confidence > 80%).

For the selected rules we extracted the spatial information of the expression domain of each gene. From ZFIN framework we get the structure names. However, ZFIN does not provide a description of the expression domains at different levels of granularity for an exhaustive coverage of the expression areas. Therefore, to complete the description of the expression domains we used the Developmental Anatomy Ontology of Zebrafish (cf. chapter 2) to derive additional spatial and functional description of the anatomical structures where the expression is observed (cf. Table 2).

Gene symbol	**Expression information**		
	Organ	**Structure**	**Functional System**
Btg2	Brain	Hindbrain, Tegmentum	Central nervous system
	Neuroblast	Neuron	Nervous System
Tbx20	Eye	Retina, Retinal ganglion Cell layer,	Visual system
	Heart	Heart	Cardiovascular system
	Brain	Hindbrain, Tegmentum	Central Nervous System
	Neuroblast	Neuron	Nervous system
Ccnb1	Eye	Eye, Optic tectum, Retina	Visual system
	Anatomical cluster	Proliferative region	-
	Pectoral fin	Pectoral fin musculature	Skeletal system
	Gill	Pharyngeal arch 3-7 skeleton	Respiratory System
Hoxa3a	Brain	Hindbrain, Rhombomere	Central nervous system
	Gill	Pharyngeal arch 3-7 skeleton	Respiratory System
	Spinal cord	Spinal cord	Nervous system

Table 2: This table shows expression information of genes of the selected rules.

In (cf. Table 2) we observed that an overlap exists between genes part of each rule. This result merits to be investigated. In this proof of principle study, we stopped at this point. In our case we used ZFIN dataset to validate and explore the PPM algorithm. Still, this result leads us to further apply the PPM algorithm on GEMS data. We integrated the

PPM algorithm within the GEMS framework so that users can run this mining algorithm on the fly.

Rule number	**ANTECEDENT**	**CONSEQUENT**
1	*myoD*	*hoxb13a*
2	*myoD*	*LysC*
3	*Fgf8*	*Shh*
4	*hoxa9a*	*Shh*
5	*sox9b*	*Shh*

Table 3: An example extracted from the result set using the PPM algorithm (support >= 30% and confidence >= 75%) on the GEMS dataset.

The patterns of gene expression are annotated with spatial variables with multi-level hierarchy. These variables could be exploited to select a dataset with common features and apply on this dataset the mining algorithm. For the rules presented here (cf. Table 3), we first generated a dataset by querying the GEMS database for patterns with a common spatial location, i.e. body and tail. Second we apply the PPM algorithm. We post-processed rules that were generated by using their annotation, i.e. temporal, functional and a spatial classification at organ and structure levels. We considered a pattern to be interesting when both its antecedent and consequent have a common spatial expression domain.

Developmental stages	24-120 hpf	36-120 hpf	18-96 hpf	10-24 hpf
Genes	*fgf8* *hoxa9a* *shh*	*myoD* *sox9a*	*LysC*	*hoxb13a*

Table 4: This table shows the temporal relationship between genes of the selected patterns.

Our experiments on the GEMS data are typically inductive. They are not applied to prove or disprove pre-existing hypotheses. Form the rules that were generated, we tried to identify spatio-temporal patterns embedded within one enclosed framework and thereby support hypothesis generation. To investigate the selected rules, we first explore the temporal characteristic of both antecedents and consequents (cf. Table 4). In rules 1 and 2, the antecedent *myoD* is expressed in early and late zebrafish development. Both consequents, i.e. *LysC* and *hoxb13a* are also expressed at early stages of development. For rules 3, 4 and 5 both antecedents and consequents have a similar temporal exhibition, i.e. at early and late zebrafish development. Second, we look at the spatial information of the expression domain of each rule. Here we explored the spatial information at different levels of granularity. We started our exploration at organ level and we finalize our exploration by looking at the anatomical structure at a finer level of granularity (cf. Table 5). Since patterns of gene expression in GEMS are also annotated with functional system information of the expression domain we used this information in our investigation. In the example below, we recognized that antecedents and consequents of rules 3, 4 and 5 have strong relationships. These relationships are seen at different levels of abstraction from body region to organ to structure to functional system. These data indicate that these genes might be strongly correlated in the morphogenesis of the posterior body in zebrafish.

This initial analysis has been realized using existing anatomical information extracted from the GEMS database. Once, a user selects a pattern of interest, a detailed analysis can start.

Gene	**Expression Domain**			**Functional System**
	Body region	**Organ**	**Structure**	
hoxa9a	Body	Fins	Mesenchyme pectoral fin bud	Locomotion
Shh	Body	Fins	Fin	Locomotion
sox9b	Body	Skeleton, Muscular and Fins	Mesenchyme pectoral fin bud and pectoral fin cartilage	Locomotion
fgf8	Body	Fins	Apical ectodermal ridge pectoral fin	Locomotion
LysC	Tail	Blood, haematopoietic tissues	Macrophages	Immune system
hoxb13a	Tail	Body axis	Tail bud	Developmental
myoD	Tail	Skeleton and Muscular	Mesenchyme fin	Locomotion

Table 5: Spatial relationships between genes of the selected patterns.

The patterns are linked to 3D images (cf. Figure 2). Requests to view 3D patterns of gene expression (3D images) are in fact 3D queries submitted to the GEMS database to visualize the expression domains in 3D. 3D patterns provide detailed spatio-temporal information of the expression domains and allow overlap discovery between genes under study (Welten et al, 2009). This 3D detailed information represents an efficient analytical approach for functional analysis at image domain. Additionally, each visualized 3D pattern is linked to external resources which provide additional dimensions for rules analysis.

The GEMS is a tool for managing and linking spatio-temporal patterns of gene expression. Here, we demonstrated that the GEMS functionality can be extended to a tool for mining patterns of gene expression. By this, we hope to create an added value to knowledge interpretation of mining results.

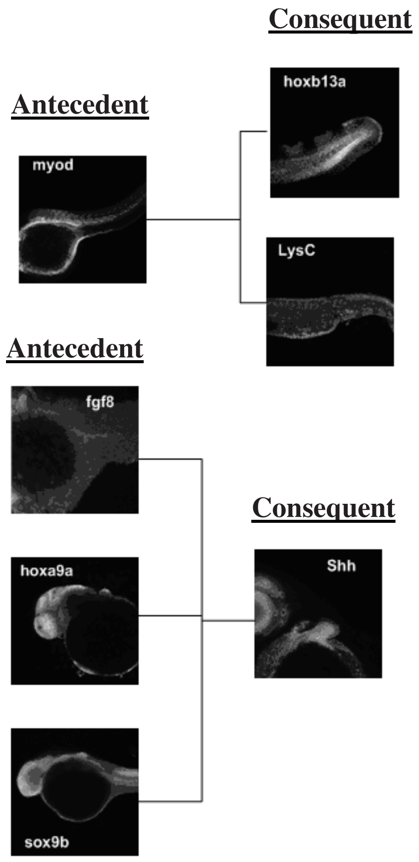

Figure 2: An example extracted from the result set of the PPM algorithm (support >=30% and confidence >= 75%) on the GEMS dataset. The first tree genes have a common expression in *tail* while the second tree contains rules with genes having a common expression in *fin* (in the body region).

6.5 Conclusions and future work

The results presented in this chapter is a proposed framework to facilitate analysis task of mining rules by improving the ability to interpret the discovered rules, evaluate their relevance and obtain insight on the discovered knowledge. We have extended our previous work (cf. chapter 4) regarding the general framework where gene expression patterns are managed using their temporal and spatial features within an integrative context. The extension includes the inclusion of mining techniques to the general framework and how to use this framework as a primary platform for mining results analysis to judge at an early stage whether a rules is interesting or not. Our experimental results are the outcome of using an association rules algorithm (PPM). Results set from this algorithm could be analysed and compared with each other. 3D patterns of gene expression (3D images) provide an advanced functional analysis of genes and spatial overlap discovery (Verbeek et al, 1999) of expression domains between genes under study. To facilitate spatial overlap discovery, direct integration of expression domains within 3D atlas models (cf. chapter 3) should be realized. This integration will allow a more advanced functional analysis in the future. Actually, the GEMS platform enables a mapping on other data resources. The patterns in the GEMS database are stored with formal and unified metadata. Therefore, the interpretation and integration of the rules within a large-scale biological network is permitted. This situation reduces the time needed to analyze the results, and prune the irrelevant rules and use interesting ones to derive new hypothesis. The preliminary results presented here, also demonstrates how generated rules may be supported by visual data representation. The researcher is able to immediately and intuitively put the discovered rule into a visual context by available gene expression 3D images.

Spatio-temporal data mining is a promising research area dedicated to the development and application of computational techniques for the analysis of spatio-temporal databases

(Mennis and Liu, 2005). Such techniques require further investigation. In this study, we started with a straightforward algorithm, i.e. PPM. Currently, we are considering other mining algorithms able to compare patterns between species and therewith including an evolutionary component. Frequent Episode Mining in Developmental Analysis is such an algorithm (FEDA, Bathoorn et al); it is based on analyzing sequences of developmental characters to find episodes. These episodes are used to determine differences between developmental sequences (Bathoorn et al, 2007). An API for FEDA should be realized to enable its execution on the fly through the GEMS which has been customized to be used as an experience bed for data mining.