Mining and Analysing Spatio-Temporal Patterns of Gene Expression in An Integrative database Framework

M. Belmamoune, D. Potikanond and Fons J. Verbeek.

Section Imaging and BioInformatics, Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands

{mounia, dpotikan,fverbeek}@liacs.nl

Summary

Mining patterns of gene expression provides a crucial approach in discovering knowledge such as finding genetic networks that underpin the embryonic development. Analysis of mining results and evaluation of their relevance in the domain remains a major concern. In this paper we describe our explorative studies in support of solutions to facilitate the analysis and interpretation of mining results. In our particular case we describe a solution that is found in the extension of the Gene Expression Management System (GEMS), i.e. an integrative framework for spatio-temporal organization of gene expression patterns of zebrafish to a framework supporting data mining, data analysis and patterns interpretation As a proof of principle, the GEMS has been equipped with data mining functionality suitable for spatio-temporal tracking, thereby generating added value to the submission of data for data mining and analysis. The analysis of the genetic networks is based on the availability of domain ontologies which dynamically provides meaning to the discovered patterns of gene expression data. Combination of data mining with the already presently available capabilities of GEMS will significantly augment current data processing and functional analysis strategies.

1 Introduction

Data mining techniques are used to identify intrinsic patterns in data, and thereby amongst other things, support generation of new hypothesis. It is recognized that the application of data mining techniques involves many tasks supported by a heterogeneous suite of tools. Typically, data analysts deal with a large number of pattern results, from which they have to retrieve the potentially interesting results and interpret what they reveal about the domain.

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 of the area under study. These requirements are however, not common to all end-users. Therefore, we propose a framework that offers data mining application and results analysis and interpretation. In this paper we present our approach that focuses on embedding mining functionality in the GEMS framework [1]. The GEMS is a system for gathering and retrieval of 3D patterns of gene expression data using domain ontologies [2]. This system has been extended to serve as an effective environment of knowledge discovery and interpretation. So, in the same framework, data mining can be applied whereas a primary analysis of the discovered rules can also be performed using domain ontologies which provides patterns with meaning and links to external resources We believe that such framework will significantly contribute to and facilitate data analysis and interpretation .

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. Importantly, in our study we focused on gene expression patterns generated through *in situ* hybridization [3] and thus include both spatial and temporal information. We have explored methodologies for data mining on our spatio-temporal data and in this paper we discuss our proof of concept methodology. In this methodology we are using an association rule mining technique to discover elements with a correlated occurrence within our gene expression dataset. Subsequently we applied pattern annotation to analyze the mining results.

Market Basket Analysis [4] is a typical and widely-used example of association rule mining. In bio-molecular research, association rules are typically applied on results of gene expression levels obtained from microarray experiments. The first step in mining microarrays is to find association rules between patterns of gene expression. The second step is to find a plausible biological interpretation of the associated patterns that are discovered. This step is the most delicate and time consuming phase in the analysis of the discovered rules since the results have to be accurately placed into context with existing biological knowledge, such as scientific literature and other biomolecular data. In our case, we work on accurate 3D patterns of gene expression that were obtained from fluorescent *in situ* hybridization experiments and annotated with standardized and structured metadata as part of the data submission to the GEMS database [1]. The way in which this information is generated and organized assures easy interpretation of mining results.

2 Methods

Association rule discovery is a data mining method that has been extensively used in many applications in order to discover associations among subsets of items from large transaction databases [4]. In order to introduce the general idea, we first formulate the conceptual framework of this method and then illustrate how we explore our data with this method. Association rules discovery methodology can be defined as follows:

2.1 Definition

1. Given a set of items $I = \{i_1, i_2, i_3, \ldots, i_n\}$ and a set of transactions $D = \{T_1, T_2, \ldots, T_m\}$, 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 implicit form:
   $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \varnothing$. The itemsets X and Y are called *antecedent* (Left-Hand-Side or LHS) and *consequent* (Right-Hand-Side or RHS) of the rule respectively.
4. Each rule is associated with its confidence and support:
   a. Confidence $(X \Rightarrow Y) = freq(X \cup Y)/freq(X)$
   b. Support $(X \Rightarrow Y) = support(X \cup Y)$,
      where $support(X \cup Y) = freq(X \cup Y)/|D|$.

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

The prototypical example to illustrate association rules is found in the domain of the supermarket [4]. Here a transaction is someone buying several items at the same time. An itemset would then be something like {cheese, beer} and an association rule is as follow: cheese $\Rightarrow$ 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 for our particular requirements. We started our gene expression mining studies with the APRIORI algorithm. We used this algorithm since it is the basic algorithm for association-rule mining. APRIORI was extensively studied and successfully applied in many problem domains [4][5]. 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 frequently and earlier in time then others. Thus considering only the occurrence count of each item (gene) may not lead to fair measurements. Therefore, we considered the Progressive Partition Miner algorithm (PPM) [6] that we applied on our set of data. The ideas of PPM algorithm is to first partition a dataset and then progressively accumulate 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.

The PPM algorithm was first validated before integration within the GEMS framework. For this validation, we used first original dataset [6] so as to get exactly the same mining results. Subsequently, we used gene expression data subset from Zebrafish Information Network, i.e. ZFIN (http://zfin.org) to further evaluate this association rules technique. We downloaded and imported ZFIN gene expression data in a local database, and then we applied the PPM algorithm on this dataset to generate associated rules. ZFIN data served as a case study to evaluate and explore the PPM algorithm against gene expression data.

2.2 Implementation

We translated the PPM algorithm into a java application that we, again, evaluated before integration into the GEMS. The GEMS has a tree tire architecture with: (1) an information side, (2) a server side and (3) a front-end (browser). Through the front-end users sent requests to the server side to be processed. Responses from the server are sent and displayed at the front-end. We defined and implemented the resources required for the interactive rule mining framework using a platform/language with java as technology support (cf. Figure 1). Through GEMS, the mining application can be executed in two different ways: as an autonomous java agent and through the user interface. In the latter case, users are able to execute the PPM mining algorithm by sending a HTTP request to execute the mining algorithm (cf. 1, in Figure 1). The application processes users' request and queries the GEMS MySQL 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. In our case we consider each transaction as a developmental stage (time point) and the items are the genes expressed at this stage. The application executes first to find the frequent 2-itemsets in the data (cf. 3, in Figure 1). From the frequent 2-itemsets the association rules between the items 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.

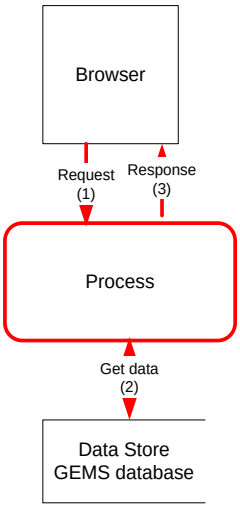

Figure 1: The process flow of the web-application to mine expression patterns.

Data stored in the GEMS concerns spatio-temporal patterns of about 100 different developmental genes in zebrafish [1]. Patterns (3D images) are obtained from whole mount fluorescent *in situ* hybridization (FISH) experiments and visualized through Confocal Laser Scanner Microscopy (CLSM). Our methodology of pattern generation enables precise information on the spatial localization of gene expression. This spatial localization significantly enhances functional analysis of gene function [3][7]. The patterns are subsequently annotated using domains ontologies and stored in the GEMS database [1][8]. We use the annotations of the patterns supported with the 3D images to post-process the mining results that we have obtained.

3 Results

ZFIN is a substantial and rich resource of (2D) gene expression data. In our ZFIN dataset, we were able to attain a large amount of rules. For PPM evaluation we limited the analysis to a small number of rules. We selected only rules in the interval of [support >40, confidence >80] (cf. Table 1). Additionally, for data analysis we limited expression information to those realized under the same experimental conditions (mRNA *in situ* hybridization) and obtained between two typical developmental stages in zebrafish ,i.e. "*prim 15*" en "*long pec*" (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 used their annotation to extract the spatial information (where each gene is expressed). From ZFIN framework we obtain the spatial anatomical annotation at structure level. 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 (DAOZ) [2]. This extended ontology provides functional description of anatomical structures and spatial information of the expression domain at different levels of granularity (cf. Table 2). For example from ZFIN we extracted the information that Btg2 is a gene that is expressed at Hindbrain and Tegmentum. For a more comprehensive description of the expression domain we used DAOZ. Therefore, we could articulate that the expression domain of Btg2 belongs to the central nervous system, is located in the brain and is exactly expressed at Hindbrain and Tegmentum.

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 illustrates expression information of genes of the selected rules.

In Table 2 we detected that the expression information resulted in a simplified analysis of rules. Therefore, we observed that an overlap exists between the expression domains of the chosen rules. This result is interesting and it will be worthy of further investigations. In this explorative study to the proof of principle, however, we paused at this point. In our experiments we used the ZFIN dataset to validate and explore the PPM algorithm. The result has allowed us to further apply the PPM algorithm on GEMS data. We integrated the PPM algorithm within the GEMS framework so that users can execute this mining algorithm on the fly while submitting new data.

The patterns of gene expression are annotated with spatial variables with a 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 in this paper (cf. Table 3) we first generated a dataset by querying the GEMS database for patterns with a common spatial location at a gross level of granularity, i.e. body and tail. Second we apply the PPM

algorithm. The generated rules were post-processed using their temporal, functional and spatial classifications at organ and structure levels.

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.

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 any pre-existing hypothesis. 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 looked 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.

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 [8]. 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.

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

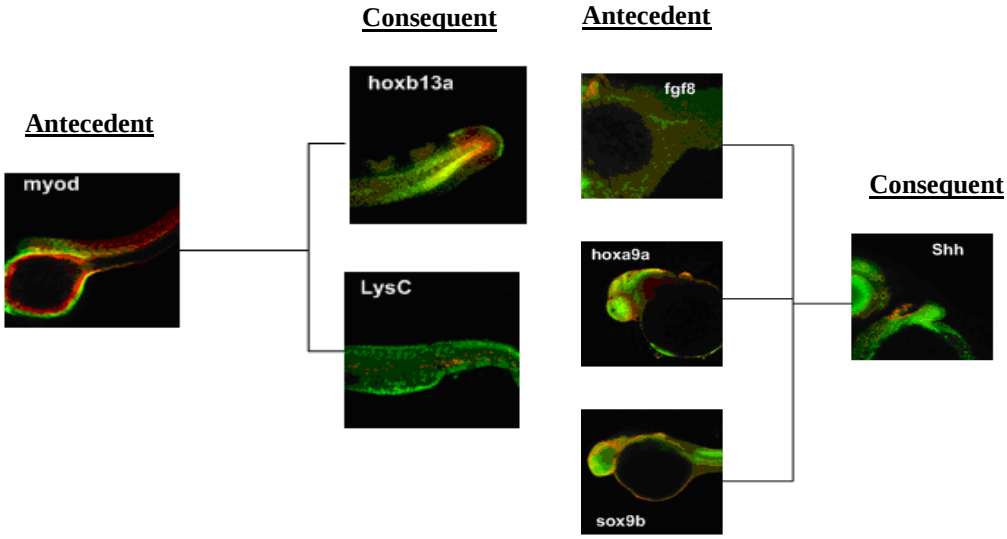

Figure 2: An example extracted from the result set of the PPM algorithm (support >=30% and confidence >= 75%) on the GEMS dataset. The first tree genes have a common expression in tail while the second tree contains rules with genes having a common expression in fin (in the body region). The depicted (thumbnail-) images of patterns of gene expression are realized from 2D projections of the dual-channels image stack that was obtained with CLSM-imaging.

4 Conclusions and future work

The results presented in this paper are part of 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 [1] 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 to analyze mining results 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). Resulting sets from this algorithm could be analyzed and compared with each other. 3D patterns of gene expression (3D images) provide an advanced functional analysis of genes and spatial overlap discovery [9] of expression domains between genes under study. To facilitate spatial overlap discovery, direct integration of expression domains within 3D atlas models [10][11][12][13][14][15] should be realized. This integration will allow a more advanced functional analysis in the future. De facto, 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 demonstrate how generated rules can be supported by a visual representation of the data. The researcher/user can immediately and intuitively put the discovered rule in a visual context given by available 3D images with patterns of gene expression.

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 [16]. Such techniques require further investigation. In this study, we started with a straightforward algorithm, i.e. PPM. Currently, we are considering other mining algorithms [17] able to compare patterns between species and therewith including an evolutionary component. Frequent Episode Mining in Developmental Analysis is such an algorithm [18][19]; it is based on analyzing sequences of developmental characters to find episodes. These episodes are used to determine differences between developmental sequences. 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.

Acknowledgements

This work was partially supported through the BioMolecular Informatics program of the Netherlands Science Organization (grant 050.50.213) and a personal grant of the Thai Government.

References

[1] M. Belmamoune and F.J. Verbeek. Data Integration for Spatio-Temporal Patterns of Gene Expression of Zebrafish development: the GEMS database. Journal of Integrative BioInformatics, 5(2):92, 2008.

[2] M. Belmamoune and F.J. Verbeek. Developmental Anatomy Ontology of Zebrafish: an Integrative semantic framework. Journal of Integrative Bioinformatics, 4(3):65, 2007. Online Journal: http://journal.imbio.de/index.php?paper_id=65.

[3] M.C.M Welten, S. De Haan, N. van den Boogert, N.J. Noordermeer, G. Lamers, H.P. Spaink, A.H. Meijer, F.J. Verbeek. ZebraFISH: Fluorescent *in situ* hybridization protocol and 3D images of gene expression patterns. Zebrafish, Vol 3. #4, pp 465 – 476, 2006

[4] R. Agrawal, T. Imielinski and A. Swami. Mining Association Rules between Sets of Items in Large Databases. Proc. of ACM SIGMOD, pages 207–216, May 1993.

[5] R. Agrawal and S. Ramakrishnan. Fast Algorithms for Mining Association Rules. Proceedings of the 20th VLDB Conference Santiago, Chile, 1994.

[6] C. Lee, M. Chen and C. Lin. Progressive Partition Miner: An Efficient Algorithm for Mining General Temporal Association Rules. IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 4, pp. 1004-1017, Jul/Aug, 2003.

[7] M.C.M. Welten, A. Sels, M.I. Van den Berg – Braak, G.E.M. Lamers, H.P. Spaink and F. J. Verbeek. Expression analysis of the genes encoding 14-3-3 gamma and tau proteins using the 3D digital atlas of zebrafish development. 2009, SUBMITTED

[8] M. Belmamoune and F.J. Verbeek. Heterogeneous Information Systems: bridging the gap of time and space. Management and retrieval of spatio-temporal Gene Expression data. In: InSCit2006 (Ed. Vicente P. Guerrero-Bote), Volume I "Current Research in Information Sciences and Technologies. Multidisciplinary approaches to global information systems", pp 53-58, 2006.

[9] F. J. Verbeek, K.A. Lawson and J.B.L. Bard. Developmental BioInformatics: linking genetic data to virtual embryos. Int.J.Dev.Biol. 43, 761-771, 1999.

[10] A. Sebastiaan Brittijn, Suzanne J. Duivesteijn, M. Belmamoune, Laura F.M.Bertens, Wilbert Bitter, Joost D. de Bruijn, Danielle L. Champagne, Edwin Cuppen, Gert Flik, Christina M. Vandenbroucke-Grauls, Richard A.J. Janssen, Ilse M.L. de Jong, Edo Ronald de Kloet, Alexander Kros, Annemarie H. Meijer, Juriaan R. Metz, Astrid M. van der Sar, Marcel J.M. Schaaf, Stefan Schulte-Merker, Herman P. Spaink, Paul P. Tak, Fons J. Verbeek, Margriet J. Vervoordeldonk, Freek J. Vonk, Frans Witte, Huipin Yuan and Michael K. Richardson (2009). Zebrafish development and regeneration: new tools for biomedical research. Int. J. Dev. Biol. 53: 835-850, 2009

[11] F. J. Verbeek. and D.P. Huijsmans. A Graphical database for 3D reconstruction supporting 4 different Geometrical Representations. In Medical Image Databases. S.T.C. Wong, ed. (Boston: Kluwer Academic Publishers), pp. 117-144, 1998.

[12] F.J. Verbeek, M.J. den Broeder, P.J. Boon, B. Buitendijk, E. Doerry, E.J. van Raaij and D.A. Zivkovic. A standard atlas of zebrafish embryonic development for projection of experimental data. Proceedings SPIE 3964, Internet Imaging: 242-252, 2000.

[13] F.J. Verbeek. Theory & Practice of 3D-reconstructions from serial sections. In Image Processing, A Practical Approach. R.A. Baldock and J. Graham, eds. (Oxford: Oxford University Press), pp. 153-195, 2000

[14] F.J. Verbeek and P.J. Boon. High Resolution 3D Reconstruction from serial sections. Microscope instrumentation, software design and its implementations. Proceedings SPIE 4621, Three Dimensional and Multi Dimensional Microscopy IX. 65-76, 2002.

[15] M.K. Richardson and F.J. Verbeek. New Directions in Comparative Embryology and the Nature of Developmental Characters. Animal Biology 53 303-311, 2003.

[16] J. Mennis and J.W. Liu. Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change. Transactions in GIS 9, 13-18, 2005.

[17] W. Meuleman, M.C. Welten, F.J. Verbeek. Construction of correlation networks with explicit time-slices using time-lagged, variable interval standard and partial correlation coefficients. Lecture Notes in Computer Science. Volume 4216, Computational Life Sciences II, pp 236-246, 2006.

[18] R. Bathoorn, and A.J.P.M. Siebes, Constructing (Almost) Phylogenetic Trees from Developmental Sequences Data. In J.-F. Boulicaut, F. Esposito, F. Giannotti & D. Pedreschi (Eds.), *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)* (pp. 500-502). Springer-Verlag, 2004.

[19] R. Bathoorn, M.C.M. Welten, A.J.P.M. Siebes, M.K. Richardson and F.J. Verbeek. Limb - fin heterochrony: a case study analysis of molecular and morphological characters using frequent episode mining. SUBMITTED.