SPATIO-TEMPORAL FRAMEWORK FOR INTEGRATIVE ANALYSIS OF ZEBRAFISH DEVELOPMENTAL STUDIES

Mounia Belmamoune

This work was carried out under a grant from N.W.O. BioMolecular Informatics research program (BMI)

Spatio-Temporal Framework for Integrative Analysis of Zebrafish developmental studies
Mounia Belmamoune.
Thesis Leiden University

ISBN 978-90-9024866-0

SPATIO-TEMPORAL FRAMEWORK FOR INTEGRATIVE ANALYSIS

OF

ZEBRAFISH DEVELOPMENTAL STUDIES

PROEFSCHRIFT

Ter verkrijgen van de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus Prof. mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 17 november 2009
klokke 15.00 uur

door
Mounia Belmamoune

geboren te Sidi Kacem, Marokko in 25 September 1972.

PROMOTIE COMMISSIE

Promotor

Prof. Dr. J. N. Kok

Co-promotor

Dr. Ir. F.J. Verbeek

Overige leden

Prof. Dr. H.P. Spaink

Prof. Dr. T. Bäck

Prof. Dr. G. Rozenberg

Prof. Dr. A. Siebes (Universiteit Utrecht)

To my parents Aicha and El Fatmi

To my brothers and sisters

To my husband

Table of Content

CHAPTER 1

GENERAL INTRODUCTION

1.1 Introduction

The specific result of vertebrate embryonic development is the progression of structures over time, from a first apparition during the developmental process to mature structures (complex organs). Throughout such developmental process genes are expressed in complex and constantly changing anatomical patterns. For anatomists, it is critical to understand how such anatomical structures function, how they change to complex shapes and which genes are involved in such changing patterns. Bioinformatics is the science that focuses on the development and application of computational methods to organize, integrate, and analyze biological-related data to facilitate the workflow for biologists. In this context we developed a spatio-temporal framework for developmental studies. A spatio-temporal reference framework of standard anatomical information and patterns of genes expression is an important tool for any experimental organism in which form and function are of interest for developmental biology. The study of anatomy is an essentially three-dimensional (3D) attempt. Therefore, to increase the value of such spatio-temporal framework, data should describe the complex relationship between tissues in three-dimensional (3D) format.

The aim of the research described in this thesis is to establish an integrative 3D spatio-temporal framework with standard anatomical information (3D digital atlas) and gene expression information (3D *in situ* patterns of marker genes) for developing zebrafish embryo; this framework has to be designed in such a way to be transposed to other model systems.

The 3D atlas of zebrafish development is a digital representation of zebrafish embryo anatomy. It provides a standardized coordinate system to analyze patterns of gene expression. The 3D atlas contains 3D digital embryos resulting from 3D reconstruction from serial sections, i.e. section images at representative stages of zebrafish development. Each of the section images is segmented in anatomical domains. Each anatomical domain

(or structure) is annotated with a graphical contour (graphical annotation). This graphical annotation enables to detect the 3D outline of the annotated structures. Furthermore, to each structure in the 3D atlas an anatomical name is assigned (textual annotation) (cf. chapter 3).

The process of gene expression refers to the event that transfers the information content of the gene into the production of a functional product, usually a protein. To be valuable for developmental studies, a gene expression information resource should be documented by its temporal (when) and spatial (where) information. The experimental conditions (how) must also be part of the documentation process for an accurate interpretation of experimental observations. We followed this workflow to manage zebrafish 3D patterns of gene expression in the Gene Expression Management System (GEMS, cf. chapter 4).

We established the GEMS that contains gene expression patterns organized and published to be readily accessed. Efforts are also ongoing in other model systems yielding to a large selection of gene expression databases such as MEPD (Henrich et al, 2005) for medaka and ZFIN (Zebrafish Information Network; http://zfin.org) for zebrafish. In the work presented here, we focused on 3D patterns of gene expression of zebrafish. This data is 3D with a spatio-temporal characteristic that provides the relation between gene expression (at a molecular level) and tissue differentiation (at an anatomical level). Such 3D representation of gene expression patterns gives molecular definitions for developmental components.

Patterns of gene expression are generated by *in situ* experiments, i.e. ZebraFISH experimental protocol (Welten et al, 2006) and the Confocal Laser Scanner Microscopy (CLSM). The patterns are 3D images basically serial optical sections carrying the spatio-temporal information of the expressed genes. The images are initially submitted as raw data to the GEMS. This enables new raw data to be readily added and integrated with other information in the database. Moreover, raw data can always be processed for

presentation according to the user's needs. Furthermore, the 3D format of the patterns enables a detailed visualization and analysis of the spatial information of the expression patterns. For valuable framework, the challenge is to map gene expression data into the atlas. A key element will be a standard anatomical nomenclature for data description in both the 3D atlas and *in situ* gene expression data.

Bioinformatics has successfully demonstrated new approaches by computationally integrating various data sets such as by using standard descriptions, e.g. ontologies to annotate collected data. Data integration is defined as the process that combines data residing at different database systems and providing users with a unified view of these data (Lenzerini, 2005). Data integration has proven to be an effective strategy to extract biological meaning from heterogeneous data sets in both developmental research and other fields. In our research we applied this principle of data integration and we developed the Developmental Anatomy Ontology of Zebrafish (DAOZ, cf. chapter 2).

The DAOZ is a key component of our information systems. It is a dictionary of anatomical terms derived from the staging series of (kimmel et al, 1993). Terms from the DAOZ are assigned to anatomical domains in the 3D atlas and are used by the GEMS as the standard nomenclature for data annotation and retrieval. This assignment represents the critical link between the atlas and the gene expression database (cf. chapter 5). The anatomical terms in the DAOZ are modeled hierarchically in different degrees of granularity. This data modeling enables complex queries to be readily performed for an intuitive data access and analysis.

Patterns of gene expression are organized in the GEMS with a standardized and structured manner and GEMS database was coupled to a 3D atlas of zebrafish development and to the DAOZ. Additionally, we added another component to our framework, i.e. the 3D visual query system (cf. chapter 5) that we developed to link the other components, i.e. the 3D atlas, GEMS and DAOZ (cf. Figure 1).

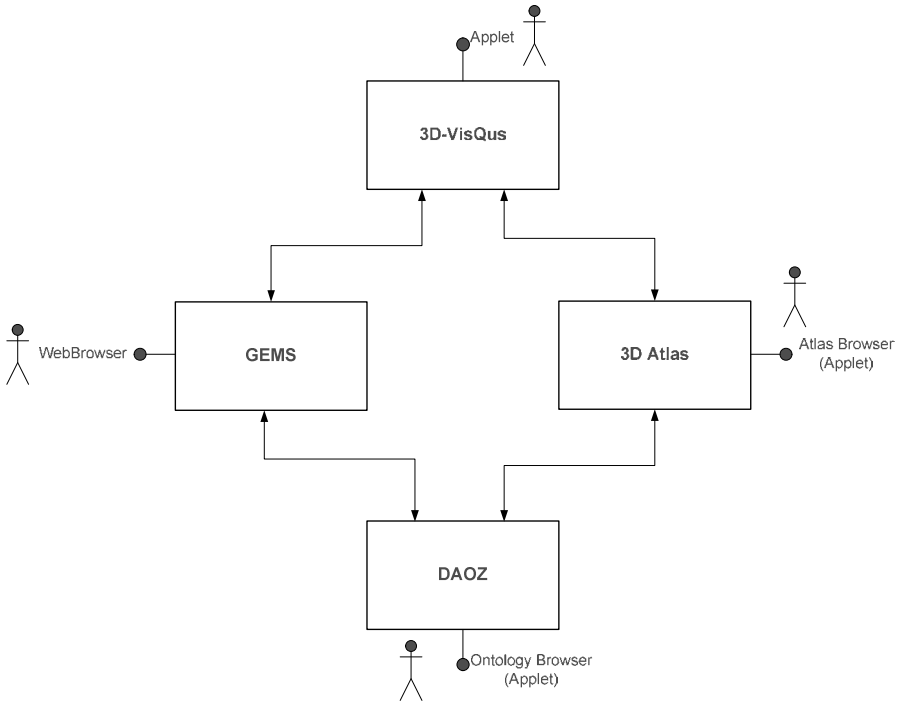

Figure 1: Diagram of the different components of our framework to study zebrafish development. Users can interact with each component through a user interface.

The role of bioinformatics is extended to uncover the wealth of biological information hidden in the mass of produced biological data and to obtain a clearer insight into the biology of organisms and to use this information to enhance the scientific benefit. In this context mining techniques were applied on gene expression data stored in the GEMS (cf. chapter 6).

With our spatio-temporal framework we introduced novel mechanisms for anatomical information storage and retrieval by using their spatio-temporal characteristics and we

make these mechanisms available to the research community in the form of novel bioinformatics tools. These tools, database systems, enable patterns of gene expression to be analyzed within a spatial and temporal context consistent with the spatial and temporal developmental concept of the organism. These resources should be seen as a tool for the developmental research community to put gene expression data into the proper biological and analytical context, so that the developmental dilemma can successively be understood.

1.2 The Zebrafish Spatio-Temporal Framework

In this part we will present the zebrafish model organism and the different components of our spatio-temporal framework for the embryonic development of zebrafish.

1.2.1. Zebrafish as a model organism

During the last decades, the zebrafish (*Danio rerio*) has become an important model organism in scientific research. Zebrafish offers a powerful combination of low cost, transparent embryo that develops rapidly outside the mother's body. Moreover, the study of zebrafish developmental genetics has proven valuable results in determining many aspects of vertebrate development. Further using this model organism promises to generate many interesting and useful data. Increasingly, it is recognized as a useful organism for human genetic and diseases modelling. The increasing use of zebrafish as a model system to study human disease has necessarily generated interest in the anatomy of this species at different developmental stages to map the many key aspects of organ morphogenesis that take place.

1.2.2 Developmental Anatomy Ontology of Zebrafish

All along of our research was the principle of data integration applied. An aspect of integration is to make databases integrated. This integration can be achieved if data in different databases are annotated with a common terminology. Ontological concepts are usually applied to provide such common terminology for data annotation in a structured way. These concepts enable therefore information sharing among different information systems.

The Developmental Anatomy Ontology of Zebrafish (DAOZ) is the anatomical ontology that we developed on behave of our project. The anatomy is modelled hierarchically from body region to organ to structure in order to fit with the different degrees of abstraction in data capture and analysis. To the anatomical and temporal concepts we introduced new concepts of spatial and functional characteristics. In addition, we used different relationships to link DAOZ concepts with each other, i.e. aggregation, composition and association relationships. These relationships provide the opportunity for more complex queries to be performed. The anatomical terminology of the DAOZ is the same as this used inside the zebrafish community. Therefore, data annotated with concept from the DAOZ ontology can be linked to each other and to other resources and more importantly to the ZFIN resources.

1.2.3 The 3D atlas of zebrafish

Core to our efforts is the 3D digital atlas of zebrafish development. The atlas is representing standard embryonic development of the zebrafish. For a number of canonical developmental stages, 3D models are generated and organized in a database system. This database contains three kinds of information: digital images, referred to as section images, graphical annotation of anatomical domains in these section images and text-based descriptions of the anatomical domains. The 3D digital atlas of zebrafish is unique because of its 3D data which is represented within a spatio-temporal context. Each

3D model is the result of a 3D reconstruction from serial sections, i.e. 2D section images. Each section image is segmented in anatomical structures that are annotated graphically and semantically.

Users can access the atlas 3D models through a web-application. This application provides a portal interface to access complex anatomical data of the 3D atlas. Users can query the 3D atlas database without a prior knowledge of the exact anatomical terms. 3D models are, on the fly, assembled and presented according to the user requested queries.

1.2.4 The gene expression data

Following the atlas, we designed and implemented the GEMS. Results from gene expression experiments sometimes redefine anatomical borders through patterns of gene expression. GEMS is a key element of our framework to study embryonic development. It is a database system for managing, linking and mining spatio-temporal patterns of gene expression in zebrafish. The patterns of gene expression are obtained from, but not restricted to, 3D images generated with the zebraFISH protocol (Welten et al, 2006) combined with the Confocal Laser Scanning Microscope (CLSM). The CLSM images show the expression domain, some surrounding tissues and the outline of the embryo. These images offer a precise approach to define gene expression domains based on reference models. These patterns of gene expression are therefore, intended to be mapped to models of the 3D digital atlas. Consequently, we applied systematic methods to manage patterns of gene expression within an integrative spatio-temporal context. The GEMS is publically accessible for data submission and inspection. Hence, integration with other resources is a key issue. The GEMS provides some level of integration with other bioinformatics resources on the Internet such as ZFIN. Moreover, integration with other model system is easier to realize.

1.3 Outline of the thesis

The work presented in this thesis is based on a number of publications in scientific journals and international conferences. Here is an overview of the chapters discussed in this thesis and their related publications.

Chapter 2 describes the Developmental Anatomy Ontology of Zebrafish. This ontology contains anatomical description of the zebrafish over time. In this chapter we will discover how the anatomical concepts have been organized as an ontology. We will also shed light on how this ontology has been translated into a database to facilitate its presentation but more importantly to facilitate its task for annotation. This ontology was initially presented in:

- Y. Bei, M. Belmamoune and F. J. Verbeek, Ontology and image semantics in multimodal imaging: submission and retrieval, Proc. of SPIE Internet Imaging VII, Vol. 6061, 60610C1 C12, 2006.

A complete description of the ontology was published in:

- M. Belmamoune and F.J. Verbeek. Developmental Anatomy Ontology of Zebrafish: an Integrative semantic framework. Journal of Integrative Bioinformatics, 4(3):65, 2007.

In Chapter 3 the 3D digital atlas of zebrafish development is presented. This chapter is partially published in:

- S.A. Brittijn, S.J. Duivesteijn, M. Belmamoune, L. F.M. Bertens, W.B., J.D. de Bruijn, D.L. Champagne, E. Cuppen, G. Flik, C.M. Vandenbroucke-Grauls, R.A.J. Janssen, I.M.L. de Jong, E.R. de Kloet, A. Kros, A.H. Meijer, J.R. Metz, A.M. van der Sar, M.J.M. Schaaf, S. Schulte-Merker, H.P. Spaink, P.P. Tak, F. J.

Verbeek, M.J. Vervoordeldonk, F.J. Vonk, F. Witte, H. Yuan and M.K. Richardson. Zebrafish development and regeneration: new tools for biomedical research. Int. J. Dev. Biol. (2009) 53: 835-850.

An advanced description of the 3D digital atlas of zebrafish development is presented in:

- M. Belmamoune, L. Bertens, D. Potikanond, R. v.d. Velde and F. J. Verbeek. The 3D digital atlas of zebrafish: 3D models visualization through the Internet. (Submitted, 2009).

In Chapter 4 we will present the Gene Expression Management System (GEMS). During embryonic development of the zebrafish, patterns of gene expression of marker genes are visualized from, but not restricted to, *in situ* hybridization experiments in combination with Confocal Laser Scanner Microscopy (CLSM). In this chapter we provide information about mechanisms of these patterns storage and retrieval. We will also give more details about the system design and implementation. The work presented here was initially published in:

- 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. 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.

The complete work has been published in:

- 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.

We will present the 3D Visual query system (3D-VisQus) in Chapter 5. This system maps standard phenotype data in the 3D digital atlas of zebrafish with genotypic data in the Gene Expression Management System. The 3D-VisQus enables 3D models of the zebrafish embryo to be viewed, browsed and queried. From a visualized element in a 3D model, a user can send a visual query to the GEMS. Questions in the kind of how this system works and how it has been designed and implemented could be further answered in Chapter 5. This chapter is based on an early publication:

- M. Belmamoune, E. Lindoorn and F. J. Verbeek. 3D-VisQuS: A 3D Visual Query System integrating semantic and geometric models. InSCit2006 (Ed. Vicente P. Guerrero-Bote), Volume II "Current Research in Information Sciences and Technologies. Multidisciplinary approaches to global information systems", pp 401-405, 2006.

To further analyze gene expression data that are present in GEMS, mining workflows have been developed. We choose for association rules techniques to investigate the mining workflow services offered by the GEMS framework. Association rules techniques have been applied to uncover possible relations between genes. Association patterns are extracted from the GEMS database and could be directly integrated with each other for a primary comparison and analysis. The uniform annotation of the gene expression data with formal ontological metadata enables cross-reference with other resources. Therefore, cross-model system comparative studies and analysis of gene expression patterns is facilitated. For more details refer to Chapter 6. This chapter is based on the following paper:

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

In Chapter 7 discussions and conclusions are presented. Also a summary in Dutch is presented.

CHAPTER 2

DEVELOPMENTAL ANATOMY ONTOLOGY OF ZEBRAFISH: AN INTEGRATIVE SEMANTIC FRAMEWORK

Based on:

M. Belmamoune and F.J. Verbeek.
Developmental Anatomy Ontology of Zebrafish: an Integrative semantic framework.
Journal of Integrative Bioinformatics, 4(3):65, 2007.

Partially published in:

Y. Bei, M. Belmamoune and F. J. Verbeek
Ontology and image semantics in multimodal imaging: submission and retrieval
Proc. of SPIE Internet Imaging VII, Vol. 6061, 60610C1 C12, 2006.

Abstract

Integration of information is quintessential to make use of the wealth of bioinformatics resources. One aspect of integration is to make databases interoperable through well annotated information. With new databases one strives to store complementary information and such results in collections of heterogeneous information systems. Concepts in these databases need to be connected and ontologies typically are providing a common terminology to share information among different resources.

Our focus of research is the zebrafish and we have developed several information systems in which ontologies are crucial. Pivot is an ontology describing the developmental anatomy, referred to as the Developmental Anatomy Ontology of Zebrafish (DAOZ). The anatomical and temporal concepts are provided by the zebrafish information network (ZFIN) and proven within the research community. We have constructed a 3D digital atlas of zebrafish development based on histology. The atlas is a series of volumetric models and in each instance every volume element is assigned to an anatomical term. Complementing the atlas we developed an information system with 3D patterns of gene expression in zebrafish development based on marker genes. The spatial and temporal annotations to these 3D images are drawn from the ontology that we have designed. In its design the DAOZ ontology is structured as a Directed Acyclic Graph (DAG). Such is required to find unique concept paths and prevent self referencing.

As we need to address the ontology in a direct manner, the DAG structure is transferred to a database. The database is used in the integration of our databases that share concepts at different levels of aggregation. In order to make sure that sufficient levels of aggregation for applications in mind are present, the original vocabulary was enriched with more relations and concepts. Both databases can now be addressed with the same unique terms and co-occurrence and co-expression of genes can be readily extracted from the databases. Integration can be further extended to the ZFIN resource and also by

including ontologies that relate to gene/gene expression (e.g. Gene Ontology). In this manner, interoperable information retrieval from heterogeneous databases can be accomplished. This greatly facilitates processing complex information and retrieving relations in the data through machine learning approaches.

2.1 Introduction

In the life sciences, data integration is one of the most challenging problems that bioinformatics is facing. In extending on new research results researchers in the life sciences have to interpret many different types of information from a variety of biological resources. Unfortunately, this information is not easy to identify and access, one of the reasons can be attributed to the semantic heterogeneity and data formats used by the underlying systems.

In this chapter, we present our approach to take up the challenge of data integration. The key is to describe and manage biological concepts into an integrated framework, leading to improved cooperation and thereby increasing scientific benefit (Baldock and Burger, 2005). In our work, we focus on the integration of data associated with the zebrafish model organism. The zebrafish (Danio rerio) is an important model organism in developmental and molecular genetics in the context of fundamental as well as disease studies. In zebrafish, experiments have produced a considerable range and huge amount of data. This fact in itself has been acknowledged by the zebrafish community and a dedicated resource, i.e. Zebrafish Information Network (ZFIN; http://zfin.org), is developed and maintained.

In the past years we have studied zebrafish development and in support of our research we have developed two important information systems. The first system is the 3D atlas of zebrafish development (3D atlas, in short); a digital atlas consisting of virtual models of *standard* zebrafish embryos at different but canonical stages of development (Verbeek et al, 1999, 2000 and 2002). The second is the Gene Expression Management System (GEMS) (Belmamoune and Verbeek, 2006). This system complements the 3D atlas by a collection of 3D patterns of gene expression of a broad range of marker genes.

The 3D atlas is the pivot in our work on developing a spatio-temporal framework for the zebrafish development; it serves as a reference for data submission and retrieval. A canonical number of developmental stages of the zebrafish are completely described as volumetric models in which every volume element is attributed to an anatomical structure. The atlas is built from serial sections portraying standard histology (Verbeek, et al, 2000 and 2002).

The GEMS is a database system for storage and retrieval of 3D spatio-temporal gene expression patterns in zebrafish including mechanisms for linking and mining. Detailed knowledge of both spatial and temporal expression patterns of genes is an important step towards analysis and understanding of complex networks governing changes during embryonic development (Meuleman et al, 2006). In our case, spatio-temporal gene expression patterns are generated through Fluorescent *In Situ* Hybridization (FISH) and whole-mount imaging (Welten et al, 2006) using the confocal laser scanning microscope (CLSM) resulting in 3D images.

For management, presentation and interoperability of the 3D images contained in the 3D atlas and GEMS, methodologies for integration need to be developed. Key is to be able come up with precise search phrases. In general, this problem is observed in an annotation phase where metadata is added to describe an object. If this, is not dealt with thoroughly, managing, mining and reasoning about information from databases will be seriously hampered. Thus, a common terminology for metadata is required. This problem is often solved with a controlled vocabulary, a series of unconnected standard concepts that is composed within a (research) community. Controlled vocabularies, however, have little to offer when it comes to reasoning by combining knowledge. It makes more sense to create agents that convey concept models with rich semantics. Ontologies are in the right position to address these issues. We have defined an approach for the annotation of our 3D images with a domain-specific ontology that implies data integration. To this end

we developed the Developmental Anatomy Ontology of the Zebrafish (DAOZ), a task-oriented ontology for annotation, retrieval and integration.

In life sciences quite a few ontologies have been developed in the model organism community. In parallel to these, the gene ontology (GO; http://www.geneontology.org/), supporting the annotation of attributes of gene products, was developed. Many of these ontologies are available from the Open Biological Ontologies resource (OBO; http://obo.sourceforge.net/) including comprehensive developmental and anatomical ontologies for many different model organisms as "Drosophila", "Arabidopsis thaliana", "Mouse" as well as an ontology for zebrafish development; i.e., the Zebrafish Anatomy Ontology (ZAO) (Sprague et al, 2006).

Our approach for handling developmental anatomy of zebrafish does not derogate the ZAO. It rather extends the ZAO with new some concepts and relationships. The DAOZ aims to provide conventions and a commonly accepted structured set of terms for annotating our research data; i.e., 3D images of *in situ* gene expression patterns. The DAOZ concepts and relationships have to supplement our 3D images with a structured annotation which is quintessential for data retrieval and mining. As a result, these annotations will enable additional comprehensive analysis of gene expression patterns during development.

Similar to ZAO, we initiated with standard anatomical vocabulary adapted from the staging series of Kimmel et al (Kimmel et al, 1995) as provided by ZFIN. The ZAO consists of two concepts types, i.e. anatomical structures and developmental stages. Anatomical structures are linked to developmental stages. In the temporal sense, each anatomical structure is defined within a time frame of start and end stage of development; this time frame records an anatomical structure as it appears and disappears during development. Anatomical structures can have relationships to each other in the ontology according to the following relationships: *is_a*, *part_of* and *develops_from*.

In the context of our work, the classes and relationships that the ZAO encapsulates are judged not sufficient to facilitate annotation, reasoning and analysis of our 3D images. The ZAO concepts and relationships limit the options for describing the inter- and intra-relationships of anatomical structures. This limitation of concepts and properties limits their use for annotation and comparative anatomical analyses. To that end, the original vocabulary has been adapted to our requirements and enriched with additional concepts and relationships. The new concepts and relationships are intended to enable descriptions of the anatomical structures in accordance with their spatial location and functional system. These concepts and their associated relations will help to structure the annotations and in that manner enabling to analyze the gene expression patterns in larger units. This is especially useful for reasoning with and mining of the data.

Similar to other ontologies, the DAOZ consists of concepts and a set of relationships. The DAOZ is organized as a directed acyclic graph (DAG); such is required to find unique concepts paths and to prevent self referencing. The nodes in the graph represent concepts and the edges joining the nodes represent relationships. Combining these relationships facilitates knowledge extraction and presentation. An important reason for using the DAOZ in annotation, apart from the consistency in the terminology for integration, is the structure in the concepts and the relations between the concepts. The relationships are intended to support retrieval of information and allow interpreting several gene expression patterns. Combining relationships also allows interpreting several gene expression patterns and obtaining information on co-localization and co-expression of genes within a common spatio-temporal framework. In this manner it can be possible to disclose “new” relations between genes.

The DAOZ incorporates terminology of anatomical structures and developmental stages identical to the ZAO. The developmental stages are the temporal concepts by which anatomical structures are organized according to appearance and disappearance during the development. In DAOZ we subsequently augment the anatomical terms conceptual

schema with additional top level concepts i.e. functional system and spatial location aspects. The concepts *functional system* and *spatial location* provide these supplementary levels of abstraction extending the data semantic and subsequently encapsulating its functional and spatial conceptual model. These concepts enable to structure anatomical terms in units using a functional system and spatial location. Searching in the ontology for concepts to annotate data is, therefore, facilitated. The annotated images are structured in the same way as their ontological metadata. This structure enables to process the 3D images in larger units which is considered useful in reasoning and mining.

To manage and use the DAOZ in a context of integration, we designed and built an ontology database. In this chapter, this database is further referred to as DAOZ. It was considered necessary to facilitate data annotation in both the 3D atlas and GEMS. Our task-oriented ontology enables interoperability and data sharing between our information system databases while cross-referencing to the ZAO is provided. Consequently, DAOZ permits integration of different information in the context of the embryonic development of the zebrafish, facilitating data analysis and knowledge extraction for presentation. The DAOZ is accessible through a user-friendly java applet.

The remaining part of this chapter is structured as follows: section 2 contains a detailed description of the adapted methods to develop the DAOZ. In section 3 conclusions and discussions are presented. Finally, section 4 describes our future work.

2.2 Methods

The major function of the DAOZ is to provide conventions and a commonly accepted structured set of terms for annotating research data; therefore, we started with the 44 staging series provided by ZFIN. This anatomical nomenclature is understandable and used by the research community and thus establishes an ideal starting point for an integrative terminology between researchers.

In this section we will describe the framework for the development of DAOZ, including conceptualization of the ontological model, relationships specification, knowledge acquisition, formal description and the subsequent choices of implementation, presentation and integration tools.

2.2.1 Conceptualization

The conceptualization phase involves identification of the key concepts in the ontology. First, we considered the anatomical structures as extracted from the staging series as our primary concepts. Second, we use temporal concepts i.e. development stages, to define anatomical terms within a range of developmental stages. For our research, however, we required an ontology that embodies more information about anatomical structures at varying degrees of granularity. Different levels of granularity enable organization of anatomical structures in units. Such organization permits integration of concepts and the objects that they describe at various levels of resolution. For this purpose, each of the anatomical terms is being evaluated and a number of paths to a certain term have been conceptualized. Two additional concepts were specified. First, specialization of functional system concepts that describe anatomical structures in relation to their functionality; e.g. ‘eye’ is described as a member of a functional system: ‘the visual system’. Second, the spatial location has been conceptualized to organize anatomical structures within a common spatial framework. This conceptualization describes the location of each anatomical domain; e.g. ‘eye’ could be described by its location in the head region. These two concepts enable to capture function and location of an anatomical structure and, as such, provide extra levels of representation for both anatomical structures as well as for our annotated images.

We further note that the scope of the ontological concepts can always be extended by adding new concepts as well as new granularities.

2.2.2 Relationships specification

We start by two hierarchical relations that were specified to describe the relationships between the various DAOZ concepts: generalization, i.e., '*is_a'* relationship and aggregation, i.e., '*part_of*' relationships (Patrick et al, 2006). The *is_a* relation specifies a generalization hierarchy between a child and its parent; e.g. 'somite 5'*is_a* 'stage of development'. With this relation a child term is linked to a broader concept. The *is_a* relationship is characterized by the fact that each child term has a transitive relationship with its parents and children, that is, properties are inherited from parents to children downstream the hierarchy, but separate properties attributed to a child term are not propagated upstream the hierarchy.

The *part_of* relationship specifies an aggregation; the idea of this relation is that individual parts are brought together into a hierarchy to construct a more generic concept. In DAOZ, we used the *part_of* relationship in two different ways. (1) "*part_of*" is used to link entities of spatial locations, functional systems or temporal concepts; in this case it does not take time constraint into consideration. For example, it always holds that 'central nervous system' is *part_of* 'nervous system'. (2) The parenthood of an anatomical structure may change over time during development (cf. Figure 1). Therefore, the *part_of* relation has been modified to incorporate temporal arguments when invoked in linking anatomical structures with each other. For example at stage '75% epiboly' (time 1) 'the presumptive brain' is *part_of* 'the ectoderm', while at stage '1 somite' (time 2) 'the presumptive brain' is *part_of* 'the presumptive central nervous system'. In both case (1) and (2) of using '*part_of*" it concerns a transitive relationship between parent and children. Such transitivity is for example expressed in a one day old zebrafish embryo where the 'retina' is *part_of* 'optic vesicle' and 'optic vesicle is *part_of* 'eye' consequently 'retina'is also *part_of* 'eye'

In order to describe anatomical structures with properties associated with spatial location, functional system and temporal conceps, we specified four associative relationships: i.e., the *located_at*, *belongs_to*, *starts_development_at* and *ends_development_at* relationships. These relationships are used to describe an anatomical term with its spatial location, functional system and developmental stages respectively. We defined each anatomical structure within a range of the appropriate developmental stages. To that end, temporal relations like *starts_development_at* and *ends_development_at* have been defined to specify time-point at which an anatomical structure appears and disappears from the process of development, respectively. Additionally, we exploit these temporal relationships to code the chronological lineage of anatomical structures during development. An anatomical structure may have several anatomical parents during its lifespan (cf. Figure 1) and therefore we coded the chronological lineage progress of each anatomical structure during its occurrence. Consequently, each anatomical term has been linked to a stage of development when it appears the first time as well as each time when its parent changes. Tracking the chronological changes over time allows following the lineage path of anatomical structures. Moreover, it enables additional reasoning about anatomical structures as well as the objects they describe.

The *part_of* relationship links two anatomical structures with each other; it attributes a specific spatial description at a fine level of granularity. We introduced the *located_at* relationship to associate anatomical terms with a spatial description at a gross level of granularity. As such each anatomical structure is associated with a spatial location concept allowing for divide and conquest strategies. For example, specifically 'retina' is *part_of* 'eye' but more generally, retina could be described by its location in head: 'retina' *located_at* 'head'. Finally, the *belongs_to* relationship is used to associate an anatomical structure with a functional system; e.g. 'retina' *belongs_to* 'visual system'.

The associative relationships also imply inheritance, so that any attribute associated with a concept describing an anatomical structure is propagated downstream by this structure;

e.g. ‘brain’ *belongs_to* ‘the central nervous system’ and ‘the central nervous system’ is *part_of* ‘the nervous system’ then ‘brain’ *belongs_to* ‘the nervous system’ too.

The associative relationships have been specified in order to describe properties associated with various anatomical concepts. Furthermore, the aggregation (*part_of*), generalization (*is_a*) and the associatives relationships are binary relationships that imply irreflexivity i.e. no term has a relationship with itself; and asymmetry i.e. if ‘retina’ is *part_of* ‘optic vesicle’ then ‘optic vesicle’ is not *part_of* ‘retina’ (cf. 2.4.2), this corresponds to a DAG.

The aggregation, generalization as well as the associative relationships aims to capture the form and the dynamic development of an anatomical structure in addition to its location and functional system.

Using DAOZ in image annotation implies that these images could later be accessed from different perspectives, amongst other things; using the anatomical structure name and also the characteristics that this structure may have: i.e. developmental stage, spatial location and functional system. Some users would use the precise term, e.g. ‘diencephalon, whereas others would use a less specific terms such as ‘brain’, ‘head’ or ‘nervous system’ to retrieve the images. Therefore, the DAOZ structure enables users to search for large data units from general concepts e.g. brain, head, and central nervous system or specifically for records from an anatomical structure name e.g. ‘diencephalon (cf. Figure 2).

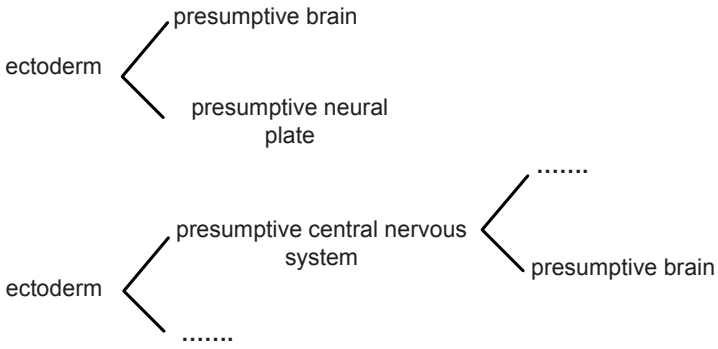

Figure 1: At '75% epiboly' is the presumptive brain part of the ectoderm, while at stage '1 somite' it becomes part of the presumptive central nervous system.

2.2.3 Knowledge acquisition

We start by the anatomical and temporal concepts as well as their relationships. The anatomical structures and stages of development nomenclature were extracted from the staging series. Information describing anatomical structures by their relationship *part_of*, *starts_development_at* and *ends_development_at*, was also extracted from the staging series. The concepts of spatial location and functional system were defined in close collaboration with domain experts. With the help of experts we established a list of attributes for the spatial locations and their relationships with anatomical structures. Concerning functional system attributes and their relationships with the anatomical structures, these have been extracted from the staging series as well as defined from both literature and domain experts. For correctness, the ontology was verified extensively.

2.2.4 Formal description

To give a more precise description of the ontology semantics, we define the concept of order (cf. 2.4.1). The concept of order is used to specify how to line up the ontology elements. Furthermore, we use 9 axioms to formalize the current representation of the DAOZ. These axioms are required as rules to check for the consistency of the ontology upon changes; as such these rules can be integrated in automated agents for ontology update (cf. 2.5).

The DAOZ consists of concepts and relationships that are organized as a DAG structure (cf. axioms 1; figure 2). In the DAG, nodes (concepts; cf. axiom 2) are linked by directed edges (relationships; cf. axiom 3). All relations imply asymmetry (cf. axioms 4) and irreflexibilty (cf. axiom 5). The *part_of* and *is_a* relationships are defined to link only attributes of the same concept type (cf. axioms 6) which means that two different attributes of different concept types could never be linked by a relationship like aggregation (*part_of*) or generalization (*is_a*). The *part_of* relationship has been modified to include time arguments in its usage to link anatomical structures concepts. (cf. axiom 7).

In a DAG each term could be linked to several parents. Therefore, each anatomical structure could be linked to other concept types thereby having more than one occurrence in the hierarchy. Anatomical structures could be associated to spatial locations, functional systems and developmental stages using the *located_at*, *belongs_to*, *starts_development_at* and *ends_development_at* relations; respectively (cf axiom 8, 9).

Definition for order in ontology

A partial order on a set S is a binary relation $< \subseteq S \times S$:

1. $\forall\, d \in S$, not $d < d$ ($<$ is irreflexive).

2. $\forall\, d_1, d_2, d_3 \in S$, if $d_1 < d_2$ and $d_2 < d_3$, then $d_1 < d_3$ (< is transitive).

Axioms underlying DAOZ

1. DAOZ is an ontology having a DAG structure.
2. A DAG *G* consists of two components: $G = SN, SE$ with *SN* is the set of nodes of *G* and *SE* its set of edges ($SE \subseteq SN \times SN$), such that for no node $n \in SN$, there are edges in *SE* forming a path from *n* to *n*.
3. *SN* consists of four mutually disjoint subsets: $SN = SA, ST, SL, SFs$. Here *SA* is the set of anatomical term concepts, *ST* is the set of temporal concepts a.k.a. developmental stages, *SL* is the set of spatial locations and *SFs* is the set of functional systems.
4. *SE* consists of 6 types of edges a.k.a. relationships, where $SE = is\_a \cup part\_of \cup belongs\_to \cup starts\_development\_at \cup ends\_development\_at \cup located\_at$.
    a. $\forall\, n_1, n_2 \in SN$ and $e \in SE$ if $n_1\, e\, n_2$, then never $n_2\, e\, n_1$. This means that all relations imply asymmetry. For example: if 'optic vesicle' is *part_of* 'eye' then never 'eye' is *part_of* 'optic vesicle'
    b. $\forall\, n \in SN$ and $e \in SE$ then never $n\, e\, n$. This means that all relations imply irreflexibility such that no concept has a relationship with itself.
5. $\forall\, n_1, n_2 \in SN_1$ with $SN_1 = SA, ST, SL$ or *SFs* (n_1 and n_2 are two concepts of the same subset) if $n_1\, e\, n_2$ with $e \in SE$ then $e \in part\_of \;\vee\; e \in is\_a$.

This means that the *part_of* and *is_a* are the only relationships linking two concepts of the same type (implying that an ordering between these exists). Consider two functional system concepts: the nervous system and the central nervous system; they only should be linked by the *part_of* relation such that 'the central nervous system' is *part_of* 'nervous system'.

6. $\forall\, n_1, n_2 \in SA$, if $n_1\, e\, n_2$ and $e \in SE \wedge e \in part\_of$ then $\exists t \in ST$ such that $n_1\, e'\, t$ with $e' \in starts\_development\_at$.

If there is a *part_of* relation between two anatomical structures we need to incorporate the time constraint since parenthood of anatomical structure may change over time during development.

7. Let SN_1; SN_2 be *SA, ST, SL* or *SFs* such that $SN_1 \neq SN_2$. $\forall\ n_1 \in SN_1$ if $\exists n_2 \in SN_2$ such that $n_1\ e\ n_2$, where $e \in SE \wedge e \notin part\_of' \wedge e \notin is\_a$.

This implies that the aggregation (*part_of*) and generalization (*is_a*) relations do not link concept types with other concept types. Thus an anatomical term can be linked to another concept type using only one of the associative relationships. For example, the only relation that links 'head' (a spatial location concept) and 'eye' (an anatomical structure) is the *located_at* relationship.

8. $\forall\ n_1 \in ST, SL$ or *SFs* and $n_2 \in SA, \neg\exists\ e \in SE$ such that $n_1\ e\ n_2$.

Any anatomical term concept can be linked to another concept type using one of the associative relations. But there is no relation that links both concepts the other way around. The relations (edges) are always directed. For example we have 'eye' is *located_at* 'head' but never 'head' is *located_at* 'eye'.

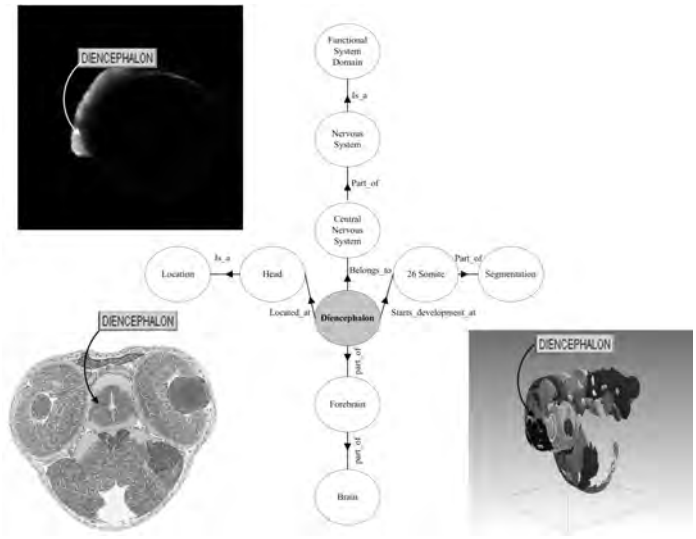

Figure 2: The diencephalon hierarchical organization to show the DAG structure of the anatomy ontology. This structure is inherited by the annotated images, e.g. top left: *msxb* gene expression pattern in a 24 hours post fertilization (hpf) zebrafish embryo, 2D projection of a 3D CLSM image. 3D model from the atlas (lower left: 2D view; lower right: 3D view of a 48 hpf. zebrafish embryo).

2.3 Implementation

To date, the most common procedure for constructing ontologies is by using tools such as DAG-Edit (http://amigo.geneontology.org/dev/java/dagedit/docs/index.html) or Protégé (http://protege.stanford.edu/). Using these tools one starts with a root term and continues adding sub-terms via connecting relationships until the ontology appears to be complete (Bard et al, 2005). In the context of our work however, we considered this an inefficient procedure. First, the DAOZ has a complex data structure with a wide range of terms and relationships, thus adding term by term will be laborious. Second, the specific aim of the DAOZ is to derive the annotation for data within other database resources. The use of the anatomy ontology in this context requires a well-designed and well-defined format that

could be easily linked to other systems and should enable complex queries to be performed to facilitate data extraction for annotation. The ontology format also should provide sufficient flexibility to permit regular updating without a need to modify the hierarchy. We therefore concluded that the anatomy ontology should be stored directly in a database, i.e. the DAOZ database.

The design of the DAOZ as DAG with a set of concepts and binary irreflexive relationships was translated to a database (cf. Figure 3). For each concept type and relationship separate tables have been designed and we assigned to each concept a unique identifier. The DAOZ database is currently implemented using the MySQL database management system.

The specific aim of the DAOZ database is to provide a common semantic framework for the annotation of our data. Therefore, it is directly linked to the 3D atlas and the GEMS to offer a common terminology for spatio-temporal data annotation in these systems. Both databases can be addressed with the same unique terms; as direct result, the 3D patterns of gene expression of the GEMS are spatially mapped onto the 3D atlas and vice versa (Belmamoune et al, 2006). Moreover, using terms from the DAOZ to annotate our biological objects means that the latter will inherit all characteristics and relationships that their annotations might embody. Henceforth, data is hierarchically organized exactly as their *ontological* metadata which is quintessential for retrieval, reasoning and mining (cf. Figure 2). Therefore, 3D images could be retrieved by anatomical structure name, as well as spatial, functional and temporal characteristics of an anatomical structure. To increase search result precision, combinatorial relationships could also be performed. For example 3D gene expression patterns annotated with DAOZ terms could be retrieved by queries in the form of “what patterns are expressed in location X” or “what patterns are expressed at time X in structures *part_of* Y”.

An ontology is never complete as knowledge progresses continuously. The organization of the DAOZ ontological concepts into a database enables updating without altering the ontology hierarchy. The actual anatomical structures of ZFIN are subject to a constant update by a consortium of researchers. We are aware that the DAOZ as well has to be validated constantly against the ZFIN nomenclature in order to improve its comprehensibility and accuracy. To this end, we developed a number of agents to maintain and update the DAOZ on the fly.

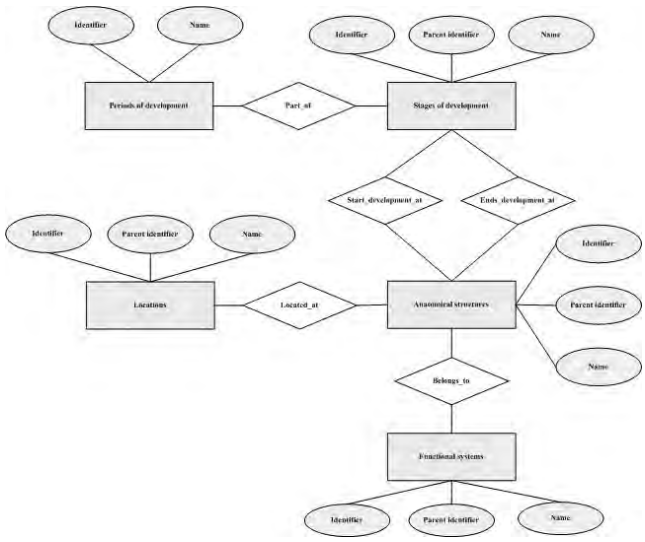

Figure 3: The entity-relationship diagram illustrates the logical structure of the DAOZ database.

2.3.1 Standalone presentation of DAOZ

In order to access the ontology, we have developed a browser: i.e., the 'AnatomyOntology'. The 'AnatomyOntology' is a java applet connected to the ontology database. The applet has been developed to enable navigation and querying anatomical terms through a pre-defined query interface (cf. Figure 4). The applet offers reasoning

possibilities; it provides users with various inference abilities to deduce implicit knowledge from the explicit represented data. The "AnatomyOntology' applet is available online (http://bio imaging.liacs.nl/liacsontology.html). In addition to the applet, on the level of database administration there is always the possibility for free-form SQL queries.

From the DAOZ database, the ontological concepts could always be represented in several common formats such as GO flat file, OBO as well as XML/RDF and OWL. To generate the DAOZ in an OBO format an additional java application, the 'OntologyGenerator', has been designed and developed. As a result, anatomical terms as present in the OBO flat file could be loaded and handled by the DAG-Edit module which offers an additional means of visualization of the data organization.

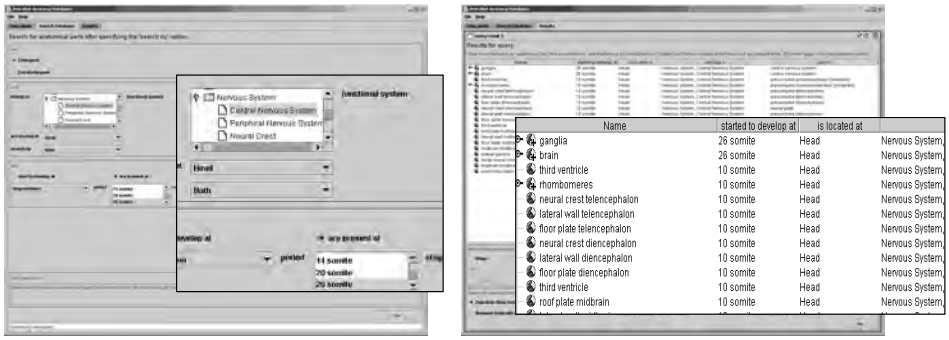

Figure 4: (Left) The applet to query the ontology database. Through this applet users are able to construct a query and submit it to the database to generate on the fly a search result. In this example we constructed the following query: 'search for all anatomical tissues present at '26 somite', belong to the central nervous system and located in head'. (Right) The result screen shows the query result with anatomical structures and their relationships.

2.3.2 Integration with other resources

The DAOZ terminology is used to annotate objects in both the 3D Atlas and the GEMS. Both databases can now be addressed with the same unique concepts and co-occurrence and co-expression of genes can be readily extracted from the databases. Another important requirement for DAOZ is to establish interoperability with other biological resources; ZFIN in particular. Anatomical terms of the DAOZ are identical to those present in ZAO; the zebrafish community ontology (ZFIN). Therefore, an object annotated with DOAZ ontological concepts can be linked straightforwardly to ZFIN which is interconnected with other database resources such as GO and the National Center for Biotechnology Information (NCBI). This means that through ZFIN, objects in our databases are integrated with others. Integration with resources such as GO and NCBI, enables our data to be presented into a large integrated research network.

GO is developed by the gene ontology consortium, and is an evolving structured and standardized vocabulary of nearly 16,000 terms in the domain of biological function (Camon et al, 2004)). GO is widely used for annotation of entries in biological-databases and in biomedical research in general.

NCBI provides an integrated approach to the use of gene and protein sequence information, the scientific literature (MEDLINE), molecular structures, and related resources, in biomedicine. Cross-references of our information systems with, but not restricted to, GO and NCBI implies integration with a wealth of bioinformatics databases leading to an increase of scientific benefit of our data.

2.4 Conclusion and Discussion

We have developed an ontology that describes the zebrafish anatomy during development based on a vocabulary established and approved by the zebrafish community. The

ontology uses several concepts and relationships for anatomical structures description which attribute numerous levels of representation. Specification of concepts and relationships has been achieved in close collaboration with experts in the field of embryology and developmental biology. As a result, the ontology provides an approved specification of domain information representing consensual agreement on concepts and relationships. Moreover, our relationships have been formally defined in order to give them uniform definitions to improve ontological consistency and to approach a maximum consistency with other ontologies; the Relation Ontology (RO) (Smith et al, 2005) especially, as it provides additional tools for relation consistency.

DAOZ is a task-oriented ontology that has been designed to annotate biological data such as 3D images of patterns of gene expression and 3D models of zebrafish embryos: i.e. the typical data in our information systems (http://bio-imaging.liacs.nl/atlasbrowserstart.html and http://bio-imaging.liacs.nl/gems/) (Bei et al, 2006). We considered it a crucial step to our efforts to implement the ontology into a well structured database that could easily be linked to other databases for data annotation. The ontology database is how we use DAOZ in applications. The structure of the ontology database is derived from the ontology DAG representation. In this database, anatomical concepts are described by unique identifiers, their anatomical, temporal, spatial and functional properties. The ontology database holds information about anatomical structures at varying degrees of granularity which enables concepts integration and descriptions at different levels of resolution; therefore complex queries could be performed against the ontological concepts to annotate data of the 3D atlas and the 3D patterns of gene expression. Moreover, powerful and complex search queries against the annotated data can be performed. The ontology is made available through a user-friendly web interface.

The DAOZ ontological concepts enable to group the annotated data in larger units. For example, the organization of spatio-temporal images with DAOZ concepts allows retrieval and integration of the relevant "*in situ*" patterns as well as obtaining information

on co-localization and co-expression of genes. This feature is very important for reasoning and mining in such data.

The DAOZ provides a common semantic framework for gene expression and phenotype annotation thus providing an integrative framework between these two types of data usually employed to study and analyze development. DAOZ improves integration and data sharing between our information systems and ZFIN as well as cross-references to other external resources, i.e. not species specific, such as GO and NCBI.

2.5 Future work

An ontology provides the conceptual framework that is used to capture knowledge in a specific domain. DAOZ concepts enable anatomical terms representation at different level of abstraction with a complex data structure. The anatomical structures are queried through a pre-defined query interface: the "AnantomyOntlogy" browser applet. This applet offers a 2D representation of the hierarchical data structure of the DAOZ. Allowing possibility of free queries as well as enabling better visualization and understanding of the ontology components and their relationships, an new improved interface to the ontology database is the route to take. Currently, we are working on the release of an interface that supports free search and allows visualization of ontological concepts and their relationships using 3D visualization. This interface is a java applet that offers a dynamic interaction with the ontology in a 3D space which will give users new insights in ontological data.

The actual ontology satisfies our requirements. However, an ontology is never complete; it can always be extended with new concepts and relationships. The RO will be extensively taken into account when new relationships will be defined in order to improve DAOZ interoperability with other ontologies. As part of the ontology ongoing development, the spatial granularity is being extended. This extension is intended to

further enrich the ontology conceptual schema. Moreover studies are in progress to realize cross-species interoperability with our ontology. A development in these ongoing studies is the recent Common Anatomy Reference Ontology (CARO) (Haendel et al, 2007). CARO is being developed to facilitate interoperability between existing anatomy ontologies for different species; this will be extremely useful in linking data between developmental model systems.

CHAPTER 3

THE 3D DIGITAL ATLAS OF ZEBRAFISH: AN INTEGRATIVE TOOL FOR ZEBRAFISH ANATOMY

Based on:

M. Belmamoune, L. Bertens, D. Potikanond, R. v.d. Velde and F. J. Verbeek.
The 3D digital atlas of zebrafish: 3D models visualization through the Internet.
(Submitted, 2009)

Partially published in:

S.A. Brittijn, S.J. Duivesteijn, M. Belmamoune, L. F.M. Bertens, W.B., J.D. de Bruijn, D.L. Champagne, E. Cuppen, G. Flik, C.M. Vandenbroucke-Grauls, R.A.J. Janssen, I.M.L. de Jong, E.R. de Kloet, A. Kros, A.H. Meijer, J.R. Metz, A.M. van der Sar, M.J.M. Schaaf, S. Schulte-Merker, H.P. Spaink, P.P. Tak, F. J. Verbeek, M.J. Vervoordeldonk, F.J. Vonk, F. Witte, H. Yuan and M.K. Richardson.
Zebrafish development and regeneration: new tools for biomedical research.
Int. J. Dev. Biol. (2009) 53: 835-850.

Abstract

We have designed and implemented a 3D digital Atlas of zebrafish development. 3D digital Atlas models have an explicit formal-ontological representation of their anatomical entities at multiple levels of granularity. This data representation is an important requirement to facilitate 3D models processing and data understanding. The Atlas is representing standard histology of the zebrafish developing embryo. Zebrafish has been established as a genetically flexible model system for investigating many different aspects of vertebrate developmental biology. It has become the focus of a major research effort into understanding the molecular and cellular events throughout the development of vertebrate embryos. The increasing use of zebrafish as a model system for developmental studies has necessarily generated interest in the anatomy of this species at different developmental stages to map the many key aspects of organ morphogenesis that take place. 3D standard anatomical resources and references that encompass the zebrafish development at early developmental stages are absent and there is therefore an urgent need for such resource to understand how different organ systems and anatomical structures develop throughout the early lifespan of this species. We have built a 3D digital Atlas of zebrafish containing a range of zebrafish 3D models. 3D models at different stages of early embryonic development are annotated with standard and formal ontological anatomical nomenclature and are made available through the internet. We have created a web-application to search, inspect and browse 3D atlas models at different levels of granularity.

3.1 Introduction

The study of anatomy during embryonic developmental is an intrinsically three-dimensional (3D) endeavour. Anatomists wish to grasp the full 3D complexity of the anatomical structures they study. A 3D shape offers an intrinsic beauty and moreover, it helps to understand how they are related to their adjacent organs and more importantly how their complex shapes are created during embryonic development. Here we present the 3D digital Atlas of zebrafish. This 3D Atlas provides the standard histology of zebrafish model organism. A 3D model consists of a variety of anatomical structures that exist at different levels of complexity or levels of biological organization (also called levels of granularity). Levels of granularity can be observed also on the level of anatomical structures functions, for instance brain (finer level of granularity) and the central nervous system (coarse level of granularity). We applied the principle of granularity to organize anatomical structures in the Developmental Anatomy Ontology of Zebrafish (DAOZ; Belmamoune and Verbeek, 2007). Furthermore, we used DAOZ ontological terms to annotate anatomical structures in the 3D Atlas models. This annotation enables Atlas data to be organized at multiple levels of granularity too. We believe that such data organization facilitates anatomical structures access; from coarse entities we get into structures at finer levels. Moreover, when anatomical entities have different levels of representation, their identity is not hampered over time than when they are treated as pure entities with one level of abstraction; for instance functions are continuous entities that preserve their identity over time. Therefore data integration becomes also stronger when entities are tracked using their attributes that navigate through different levels and over time. The work presented here is an extension of the already published 3D digital Atlas of zebrafish (Verbeek el al, 2000, 2002). 3D models of zebrafish at early developmental stages were produced and published to allow users to learn more about anatomy of the developing zebrafish embryo. In its new version, 3D models of the Atlas were evaluated and annotated with standard nomenclature from the

DAOZ and reorganized in an object oriented database. Through internet tools users are able to interact with the Atlas database and perform queries against the 3D models. A 3D Atlas model is an entity that through these internet tools could be accessed and browsed as a whole as well as through its instances. For example users could view a 3D model of 36 hours post fertilization (hpf) as a whole or in sub-structures by composing a search query such as: “select all anatomical structures from a 36 hpf embryo that belong to the central nervous system and located in head”.

Zebrafish (Danio rerio) has emerged as a useful model system to study vertebrate development. The zebrafish model system has the considerable advantage of holding complex developmental systems not present in other model invertebrates such as C. elegans and Drosophila (Lieschke, 2001). Prior to our 3D digital Atlas, no detailed 3D anatomical reference for the early zebrafish embryo was available. The stages of zebrafish development have been intensively studied over the last decades, and an embryological staging series (Kimmel et al, 1995) has been provided including key events in embryo development. However, a detailed 3D documentation of zebrafish development describing the whole anatomy of the embryo was not available. Previous detailed descriptive studies of zebrafish development were presented in 2D or limited to a particular functional system (Isogai et al, 2001). A holistic understanding of the anatomy of an organism is critical to dissecting the development and function of different organs and tissues in space and time.

Imaging methodologies such Optical Projection Tomography (OPT) and Magnetic Resonance Imaging (MRI; Kabli et al, 2006) were developed to rapidly and easily generate digital data on the internal structure of tissues without the necessity for cutting the sample, although even today the results are most commonly viewed in two dimensions, as virtual sections. This is the approach used by FishNet project that uses OPT to generate 3D models of larval zebrafish (Bryson-Richardson et al, 2007). With the 3D Atlas of zebrafish we attempt to build a 3D reference framework with real standard

anatomical structures and at high resolution. Moreover, this 3D framework should be ready for interoperability with other resources for an improved developmental study.

The 3D models of our digital Atlas are the result of 3D reconstruction from serial physical sections (Verbeek et al, 1998; Brune et al, 1999; Weninger et al, 2002). Each anatomical domain in a 2D section is outlined by a closed contour and annotated by a standard nomenclature. With this imaging method we get high quality models providing detailed view of the anatomical structures in the plan of section images. This technique is preferred over other non-destructive methods such as OPT and MRI. The most obvious advantage of using histology is the high level of detail which can be achieved. Furthermore it enables the use of staining methods, thereby enhancing the contrast between tissue types, which makes images annotation (next section) easier. This process of 3D models generation might be qualified as time-consuming however each embryo, i.e. 3D model could be considered as a standard model for developmental biology and therefore justifies the amount of effort involved in its production.

The effectiveness of any model organism is restricted by the availability of accurate anatomical information for that model organism. With the 3D Atlas we strive to broaden the understanding of the biological development of the zebrafish and provide stakeholders with information on the scope of the size and shape of anatomical domains so as to enable an anatomical structure to be compared with its well described presumed "standard". For zebrafish developmental studies gene expression patterns are produced at different stages of development. To understand this wealth of molecular data the 3D Atlas serves as a spatial and temporal mapping system for gene expression data submission and retrieval.

The 3D digital Atlas of zebrafish consists of a number of canonical stages of the zebrafish embryo and we have chosen to report on the early stages that have been completed; these are 24, 36, 48 and 72 hpf. These stages correspond to those of the

staging series of the early work of kimmel et al. For each of these developmental stages an acquisition database was produced. 3D models, i.e. 3D images in this database were graphically and semantically annotated. The graphical annotation is realized by means of closed contours around anatomical domains while the semantic annotation is given by conveying to each anatomical domain its presumed anatomical name.

The 3D Atlas is intended to be used as a standard framework to understand development of the zebrafish embryo. Development is a spatial and temporal event; therefore the 3D models of the Atlas need a spatio-temporal description. The spatial, temporal as well as function information in the 3D models is provided from the DAOZ (cf. chapter 2).

At present, biological research information is preferred in a digital computer readable form so that this information can be shared and linked to other digital resources. This is in particular important for anatomical Atlases of model systems. We have, therefore made our 3D digital Atlas available through internet. Atlas images and annotations are organized in a database system. Furthermore, we developed a web-application, i.e. ZFAtlasServer (http://bio-maging.liacs.nl/ZFAtlasServer) that enables users to render 3D models on the fly. More importantly it allows users to search in a 3D model for specific instances or anatomical entities that are subsequently displayed on the fly, in 2D and 3D formats.

The remaining part of this work is structured as follows: section 2 contains a detailed description of the adapted methods to acquire and construct 3D models in the Atlas. In section 3 we describe how the 3D model are annotated and in section 4 we presents the way 3D models are managed in an object oriented database. Furthermore, in section 5 we give a detailed description of our internet tools for 3D models retrieval and visualization. Section 6 is dedicated to present our results and discussions. While our future works are discussed in section 7.

3.2 3D Models Acquisition

Here we present the model's acquisition. So far we produced a number of embryos along the time axis at the following developmental stages 24, 36, 48 and 72 hpf (staging was based on Kimmel et al., 1995). The models presented here are acquired on normal and high resolutions.

3.2.1 Imaging methodology

Our starting point is a series of serial sections (Verbeek et al, 1995, Verbeek, 2000 and Verbeek el al, 2000). From these sections, section images were acquired with our dedicated acquisition station (Verbeek & Boon, 2002); Zeiss microscope equipped with CCD camera (JAI-Vision) and a xy-stage (Marzhauser) controlled by a MAC 4000 (Marzhauser). The stage controller was connected to the computer through the serial interface (RS 232) while the CCD camera was connected to the computer via a PC Vision frame grabber. Our acquisition software, 3Dacq, vs. 2.0 (Verbeek et al., 1998; Verbeek and Boon, 2002) controlled both the XY stage and the camera/frame grabber and thereby also the image capturing as well as the positioning of the region of interest in the field of view of the microscope. The PCVision frame grabber facilitated video overlay and this was utilized in the sequential acquisition of the section images. While the previous image section was kept in overlay, the next section was acquired and aligned with the overlay. In this manner an accurate registration of the consecutive initial section images was obtained. The logistics of the acquisition process were assembled in an acquisition database that was transposed to a data set suitable for annotation.

3.2.2 Normal Resolution

The images of normal resolution were acquired using standard camera output. The whole process of producing an initial set of section images at standard resolution for the 3D modelling was very well documented in a database. This documentation is required for high resolution acquisition (as described in the next section). After processing, all section images were stored in Portable Network Graphics format (PNG; http://www.w3.org/Graphics/PNG/).

3.2.3 High Resolution

At high resolution models histology could be seen in more details than at normal resolution. The XY position and angle φ of the camera during standard resolution acquisition has been stored in the acquisition database for each section image and were used to obtain a second set of images at a higher magnification. High resolution series were produced for the developmental stages of 36 and 72 hpf embryos. In the high resolution series each image was constructed from multiple tiles. An overlap of 32 pixels between the tiles was used to be able to assemble the tiles into a complete image section after acquisition. The obtained tiles were combined into image sections and preprocessed using the methods described in Verbeek and Boon, 2002. The images acquisition was accomplished using the normal resolution image stacks. The image stack was annotated and visualized in the same way as at normal resolution.

3.3 3D Models Annotation

Image stacks for several developmental stages have been produced and annotated. Each image has been segmented into multiple anatomical areas (or structures). Each anatomical structure is outlined by a closed contour; this makes it an anatomical domain.

Furthermore, we assigned an anatomical name to each domain. This anatomical nomenclature is extracted from the DAOZ database.

The annotation and subsequent visualization of the resulting 3D models was realized with a software suite that we developed for generating 3D models out of plan parallel sections, i.e. the TDR-3Dbase (Verbeek et al, 1995 and 2000) (cf. Figure 1). This phase of the 3D reconstruction starts with transfer of the acquisition database with the section images and other relevant information to TDR-3Dbase; together with the annotated structures this transposed to a (XML) database dedicated to the 3D reconstruction.

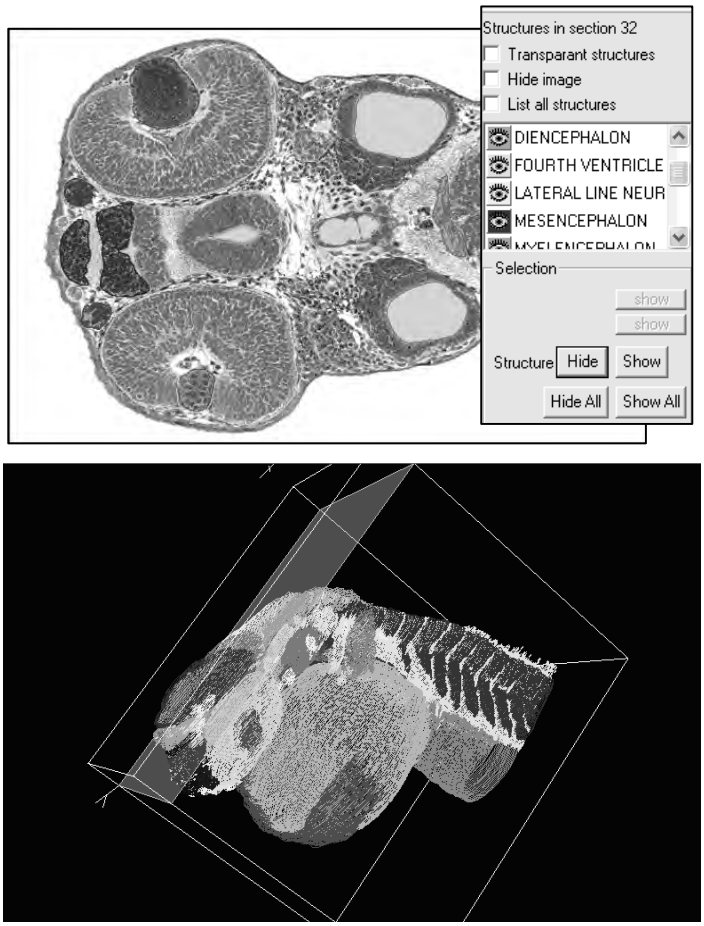

Figure 1: (Above) shows a section image that has been segmented in multiple anatomical domains (closed contours) and to each anatomical domain a color and an anatomical name are assigned, these are given in the form of an anatomical concepts list. (Below) illustrates the 3D contour model; it is the result of 3D reconstruction using contour elements.

The 3D Atlas consists, therefore, of two parts; one is a textual description of the anatomical structures and the other is a graphical annotation realized by segmenting each section image into multiple domains. The first part is stored in an object oriented database while the second part is stored on the server file system. Both parts are accessible through the internet for 3D models visualization.

3.3.1 Graphical annotation

To enable a wide range of users to understand what anatomical information is present in a section image a graphical annotation is required. This form of annotation indicates a domain in the image (area or volume) in which an anatomical concept is observed. In each of the section images anatomical structures were traced using a WACOM LCD tablet (PL series, WACOM, Europe). To that end, each anatomical structure was specifically named and attributed a color label.

3.3.2 Textual annotation

A textual annotation is realized by attaching a name to each anatomical domain in every section image. Anatomical domains are annotated with anatomical names from the DAOZ. The annotation of the Atlas 3D models with a domain-specific ontology implies its data integration with a broad range of resources. Additionally, anatomical structures in the DAOZ are organized at different levels of abstraction. To each anatomical term a number of path to that term have been conceptualized such as its spatial location, functional system and stages of development. This organization allows structures to be grouped in units at different levels of granularity. This data organization is inherited by the annotated anatomical domains. It provides therefore a novel mechanism to retrieve and group anatomical domains at different levels of details (from gross to finer levels of granularity) for example from functional system (central nervous system) to structure (brain) to sub-structure (midbrain). Moreover, a wide range of users can readily search

the Atlas. They are able to search in a 3D model for detailed information (instances) using coarse anatomical entities. Anatomical entities in an Atlas 3D model are described at multiple levels of granularity which enables their identity to remain maintained through time. This makes of the 3D digital Atlas a stronger and more persistent data integration framework where anatomical structures and attributes traverse multiple levels.

3.4 3D models Pre-processing and Management

For reasons of flexibility, we choose XML to be the native format of our data. As XML is involved in the whole processing pipe, it is relatively easy to add entries and/or attributes to the XML files without having to rewrite all software. This scalability feature is indispensable for a project which is subject to adaptation and update so as new insights are added to constantly improve the quality of the data. The Atlas data comes from two resources; one being the image acquisition software and the other being the TDR-3Dbase software package (Verbeek et al, 2002). Using the TDR-3Dbase 3D reconstruction from serial sections is realized and for each 3D model an XML file is generated.

An XML file includes entries for images and anatomical terms. Each anatomical structure is attached to an encoded file that stores contour information of its corresponding anatomical domain. These entries are extracted from the original XML files and stored in the Atlas database; PostgreSQL is the Database Management System (DBMS) that we are using. This process is realized using a parser software that takes an XML file as input and organizes its content into the Atlas database. Organizing data in an object oriented database facilitate data access and retrieval. In other words, search queries could be performed against the database to choose models or sub-models of interest.

3.5 Data Delivery: An interface for the Atlas database

Having created models covering the important early stages of embryonic development we designed a web-application, ZFAtlasServer, to allow easy online access to the full Atlas dataset. From an earlier application we have elaborated and produced an extension for the web-aplication: AtlasBrowser (Verbeek et al, 2002). The AtlasBrowser is a java applet. It allows data display and browsing in 2D and 3D formats. The ZFAtlasServer extended the AtlasBrowser by enabling 3D models to be queried and generates query results on the fly.

ZFAtlasServer is a three-tier application with three layers (cf. Figure 2). (1) The presentation layer or the Graphical User Interface (GUI). Through the GUI users send search requests and get the results prompted in an applet. (2) The application logic layer. This layer processes that services data requests between the user and the databases. (3) The databases or the data layer is where data is managed.

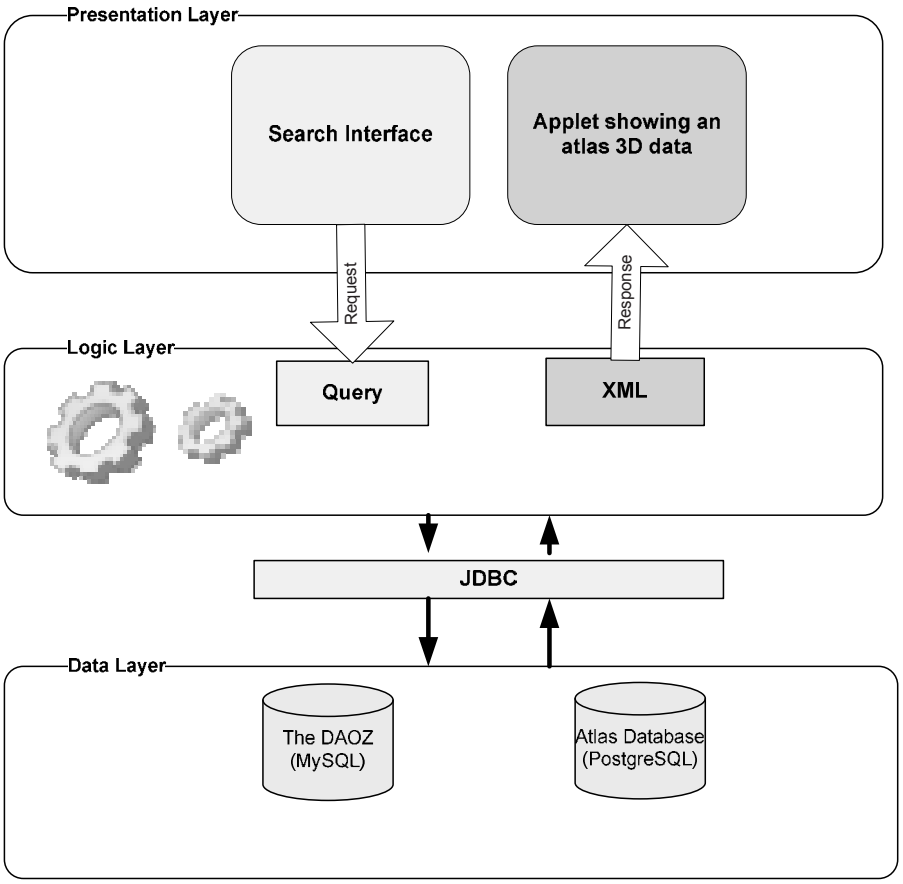

Figure 2: The user-server architecture of the web application: ZFAtlasBrowser.

Through a search interface, a search query could be performed. A query is composed using conceptual paths to the annotated anatomical domains, e.g. “select in the 36 hpf embryo all anatomical domains that belong to the nervous system and located in head”. This query is executed against both the DAOZ and the Atlas databases. A list of anatomical structures is returned. This list is further used by the services layer to generate

an XML file with all required elements as entries to the images set and their textual and graphical annotations. This XML file is passed to the java applet application. Through the applet, a volume rendering of the retrieved anatomical domains is prompted to the user (cf. Figure 3). The user is able to navigate through the data in 2D or 3D.

We have generated two versions of 3D models: at normal and high resolutions. Users can choose to query and browse the normal resolution models before selecting a higher resolution copy to examine in more detail. In this way, the volume data can be easily browsed in an intuitive manner on the wide range of operating systems and internet connections used by researchers. The ZFAtlasServer includes the following features:

1 Display a whole 3D model
2 Search in a 3D model for specific anatomical structures
3 Display the search results
4 Display 2D section images
5 Display contours present in each section image
6 List of anatomical structures nomenclature
7 Graphical legend of the anatomical structures
8 Tabs for 2D and 3D visualizations
9 Select/deselect anatomical structures options
10 Features to navigate, zoom in/out and to manipulate the images opacities.

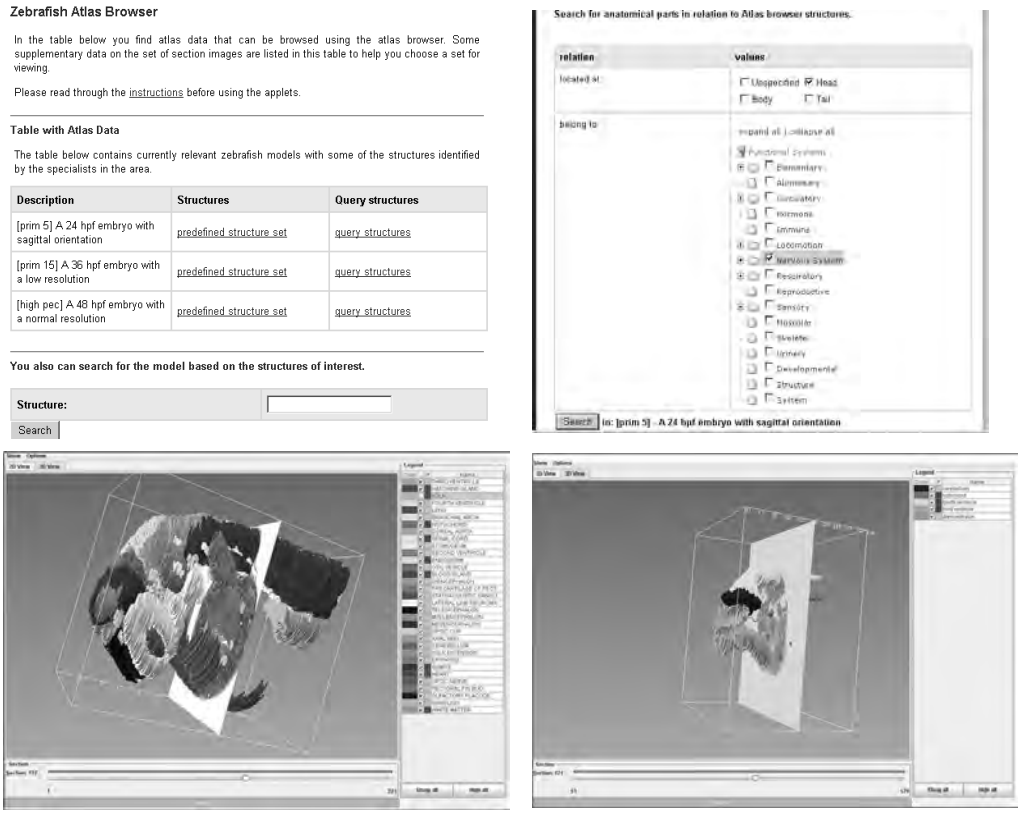

Figure 3: (Left, above) the first screen where users could directly start the java applet to view a complete 3D model. (Right, above) The query interface is where users are able to compose a search query using entities from the DAOZ ontology. (Left, below) A whole 3D model of a 36 hpf embryo. (Right, below) A 3D sub-model (the central nervous system in the head region) of the 36 hpf embryo; it is the result of the following query: "Select all structures present in the 36 hpf embryo, belong to the central nervous system and located at in head".

The ZFAtlasServer supports other geometrical representations. We developed another application to display surface models. The TDR-3Dbase uses contour information to generate 3D object models. For the object models, a triangulation algorithm is used, i.e. Boissonnat algorithm (Boissonnat, 1984; Verbeek et al, 1993) that enables acquisition of

triangular surface tiles between consecutive contours in section images. The triangulation operation is used to create visually attractive 3D surface models (Verbeek and Huijsmans, 1998). Triangulated objects are created for each anatomical domain. Entities are added into the database to describe each anatomical domain with its surface model information (cf. Figure 4). To enable surface models visualization, the Atlas web-application has been extended with the object viewer tool. This extension is needed to offer an advanced form of data visualization through the internet.

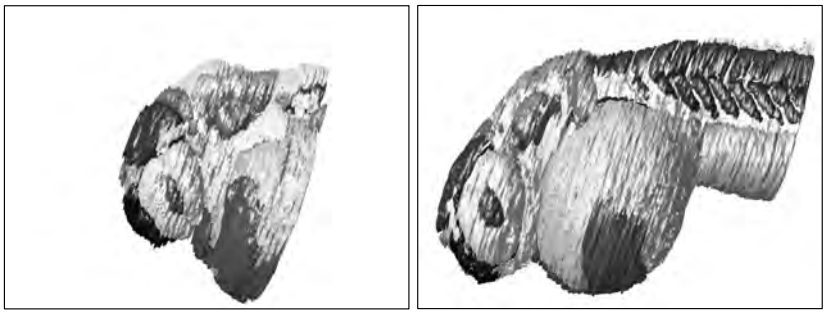

Figure 4: (Left) 3D surface model of the hear region of a 36 hpf embryo. (Right) The 3D surface model of a whole 48 hpf embryo.

3.6 Results and Discussion

We collected numerous samples and we made the best ones available online: at 24, 36, 48 and 72 hpf. These samples (or 3D models) represent important embryonic development stages of the zebrafish. We presented here the technology that we adopted to generate the 3D Atlas models. The models are the result of 3D reconstruction from physical serial sections. This method allows detailed cells observation and is therefore still preferred over other non-destructive methods such as OPT and MRI. The overall advantage of physical sections is that one can observe cells at different resolutions and study the cells in the context of the histological tissues and textures in the object. Each embryo, i.e. 3D

model generated with this powerful method could be used as a standard by the scientific community.

Section images of a 3D model are annotated both textually and graphically. Textual annotation helps to understand what is in the image while it does not help to indicate where in the image a certain concept is observed. Words help to trace images, not domains in the images. Sometimes the annotation is very obvious; in other cases a graphic clue is added to help to locate a concept in an image. The simplest form of a graphic aid is to add an arrow or a line pointing to an area in which a concept is observed. This is not unambiguous in all cases. It can be easily implemented by using Scalable Vector Graphics (SVG; http://www.w3.org/Graphics/SVG/) which provides XML structures to do this. A more precise way of using a graphic aid is to apply a true graphical annotation by indicating in an image a domain, i.e. a group of pixels with similar characteristic. In our case, we segmented each section image by drawing closed contours around anatomical structures sharing similar anatomical functionality. Coordinates of each contour element (or graphical annotation) are saved as part of the annotation in a contour file.

The graphical annotation is completed by assigning to each anatomical domain an anatomical nomenclature. We used anatomical concepts from our formal ontology, i.e. DAOZ. These concepts are organized at different levels of representation. Therefore, a gross level granularity could be used to access anatomical information at a finer level of granularity. Annotated anatomical domains could also be accessed at different levels of granularity and grouped in units according to users' queries. Therefore, from one sample embryo in the database, multiple instances could be generated on the fly according to users search desires. Additionally, viewing an instance instead of the whole embryo reduces enormously the load time since we have less data transfer.

The DAOZ anatomical structures nomenclature is the same as this known and used inside the zebrafish research community. DAOZ concepts enable the 3D Atlas data to be integrated with other resources such as ZFIN; objects in the Atlas could be integrated with others resources which enable our data to be presented into a large integrated research network. Anatomical domains are annotated with continuous entities such as functional systems and stages of development which means that they preserve their identity over time even when lower level anatomical nomenclature are changed. This feature facilitates Atlas data access and integration to other model systems.

Annotated models are organized in an object oriented database, i.e. the Atlas database. This database includes information of each 3D model. This information is related to a model anatomical structures nomenclatures, graphical annotation as well as pointers to their set of section images. A query interface has been setup to access objects in this database. From this interface, straightforward queries could be performed using ontological characteristics of the anatomical structures. Users are not required to have a deep knowledge of the anatomy to interact with the Atlas. Therefore, we trust that a wide range of users can made use of our 3D Atlas.

When a search query is sent to the database an XML file of the query model is generated containing query results. The XML file is used as an input for the applet application. This application enables models visualization and manipulation on the fly. In addition, a model could be freely manipulated, such as structures could be isolated to be studied individually or in the context of their neighbors.

The zebrafish 3D Atlas provides a novel and valuable tool for researchers studying zebrafish embryonic development and can be applied to a range of research areas, including the identification of abnormal anatomical patterning in transgenic lines. In our case, the 3D Atlas of zebrafish already serves as a model for 3D gene expression information submission and retrieval (Belmamoune et al, 2006, 2008). The techniques we

have developed and employed to acquire, manage and present the data have been successfully applicable to many other model systems and anatomical structures (De Jong et al, 2005; Welten et al, 2005 and Bertens et al, in preparation).

3.7 Future work

The Atlas presents a novel, accessible, intuitive approach for studying zebrafish anatomy that will facilitate analysis of embryo morphology. We also expect that it will be an excellent reference tool for a broad range of the scientific community and be more useful as an educational tool. By propagating the Atlas use in the community, we will get feedback from users about any inconsistencies in data annotation and how features of the datasets and software can be improved. Work has been realized to improve mapping the Atlas models to our 3D spatio-temporal patterns of gene expression (cf. chapter 5). Moreover, mapping the zebrafish Atlas to Atlases such as the Edinburgh Mouse Atlas Project (http://genex.hgu.mrc.ac.uk) and the Edinburgh Mouse Gene Expression Atlas (http://genex.hgu.mrc.ac.uk) is under consideration. As more data are added to the Atlas database we will continue to improve our tools to access this data set to disseminate the use of this valuable 3D resource.

CHAPTER 4

DATA INTEGRATION FOR SPATIO-TEMPORAL PATTERNS OF GENE EXPRESSION OF ZEBRAFISH DEVELOPMENT: THE GEMS DATABASE

Based on:

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.

Partially published in:

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.
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.

Abstract

The Gene Expression Management System (GEMS) is a database system for patterns of gene expression. These patterns result from systematic whole-mount fluorescent *in situ* hybridization studies on zebrafish embryos. GEMS is an integrative platform that addresses one of the important challenges of developmental biology: how to integrate genetic data that underpin morphological changes during embryogenesis. Our motivation to build this system was by the need to be able to organize and compare multiple patterns of gene expression at tissue level. Integration with other developmental and biomolecular databases will further support our understanding of development. The GEMS operates in concert with a database containing a digital atlas of zebrafish embryo; this digital atlas of zebrafish development has been conceived prior to the expansion of the GEMS. The atlas contains 3D volume models of canonical stages of zebrafish development in which in each volume model element is annotated with an anatomical term. These terms are extracted from a formal anatomical ontology, i.e. the Developmental Anatomy Ontology of Zebrafish (DAOZ). In the GEMS, anatomical terms from this ontology together with terms from the Gene Ontology (GO) are also used to annotate patterns of gene expression and in this manner providing mechanisms for integration and retrieval. The annotations are the glue for integration of patterns of gene expression in GEMS as well as in other biomolecular databases. On the one hand, zebrafish anatomy terminology allows gene expression data within GEMS to be integrated with phenotypical data in the 3D atlas of zebrafish development. On the other hand, GO terms extend GEMS expression patterns integration to a wide range of bioinformatics resources.

4.1 Introduction

Patterns of gene expression are studied in all major animal model systems in a systematic manner and the data resulting from gene expression studies are stored into a range of model organism databases. The major databases are FlyBase (Grumbling et al, 2006) for *Drosophila*, MEPD (Henrich et al, 2005) for medaka, GXD (Smith et al, 2007) and EMAGE (Christiansen et al, 2006) for mouse and ZFIN (http://zfin.org) for zebrafish. In our work we focus on patterns of gene expression resulting from developmental processes involving both molecular (transcription) and morphological (genotype) data. This data clearly have a spatio-temporal signature.

In this chapter we present the design and implementation of a repository for patterns of gene expression in zebrafish derived from 3D images, i.e. the Gene Expression Management System (GEMS; http://bio-imaging.liacs.nl/gems/). This system aims to be an integrative database for spatio-temporal patterns of gene expression with other bio-molecular databases, crucially important for an efficient use and exchange of gene expression resources. For zebrafish a central repository for true 3D patterns of gene expression is needed. We therefore investigated integration of 3D patterns of gene expression with bimolecular databases; first to link image information with genomic information and second to study interoperability between genomics model systems.

Zebrafish is an important model organism used in molecular genetics and developmental biology; it serves as a model for understanding normal vertebrate development as well as dissecting the mechanisms underlying human diseases. As a vertebrate model, zebrafish has many advantages: small size, ease of culture and transparent embryos. Moreover, many aspects of vertebrate development can be compared with zebrafish.

In our work we focus on 3D spatio-temporal patterns of gene expression in zebrafish which are generated through Fluorescent *in situ* Hybridization (FISH). We have

specifically improved and adapted a FISH-protocol for the imaging of whole mount zebrafish embryos using the Confocal Laser Scanner Microscope (CLSM): i.e. ZebraFISH (Welten et al, 2006). Each CLSM image is a 3D multi-channel image taken from a whole mount specimen containing the outline of the embryo and the pattern of gene expression in separate channels. However, the patterns of gene expression are not restricted to the zebraFISH protocol. It is the intention to accommodate different kinds of protocols so that patterns of gene expression can either be the result of FISH, or transgenic lines (GFP-like) or immunohistochemisty (product related). With respect to 3D CLSM images resulting from zebraFISH, the GEMS repository realizes storage, retrieval and mining of these patterns of gene expression, in coherence with their spatial and temporal characteristics. For this particular domain the GEMS aims complementing the comprehensive Zebrafish Information Network (ZFIN; http://zfin.org) as a platform integrating zebrafish 3D spatio-temporal patterns of gene expression. Data annotation is a crucial aspect of the GEMS and this is accomplished through the integration of two domain ontologies: the Developmental Anatomy Ontology of Zebrafish (DAOZ; http://bio- imaging.liacs.nl/liacsontology.html) and the Gene Ontology (GO; http://geneontology.org).

One of the important challenges in developmental and molecular genetics is to determine how genes interact to control biological processes. In developmental biology, this task is even more challenging since one attempts to understand the complex genetic process underlying development (Meuleman et al and Gilbert, 2006). Identifying both temporal and spatial aspects of gene expression in development is a critical initial step to prepare the groundwork for additional functional analysis of genes. To this respect, the microarray technique is one of the major experimental breakthroughs enabling high throughput measurement and analysis of the expression patterns of (tens of) thousands of genes simultaneously (Basset et al, 1999). In case of model organisms, the analysis of whole-specimen microarray gene expression data does not give a sufficiently specific

spatial profile. In multi-cellular organisms such as the zebrafish, gene expression influences and/or directs the developmental status of a cell or group of cells. Whole-specimen microarray analysis can therefore not fully document the spatio-temporal relations. Whole mount *in situ* hybridization can be used to obtain such information. 2D images are most commonly used and easier to acquire, using standard *in situ* hybridization. However, this may not be sufficient in describing complex phenomena as development. 3D images are used to obtain a comprehensive description of the spatio-temporal relations and have the additional advantage that internal anatomy can included in the visualization. For a range of marker genes zebraFISH *in situ's* have been done and the resulting 3D images (cf. Figure 1) are stored in the GEMS database. The 3D patterns of gene expression are available as "raw" image data and in some cases also 3D graphical models (Welten et al, submitted) are extracted from the 3D images and are made available.

First of all, GEMS is a database that is set out to manage data in a broader context. The key to success of any database is the existence of an appropriate semantic framework for its data storage and retrieval. 3D patterns provide gene expression information within a spatio-temporal context therefore, this kind of data when stored in the database, require both gene and spatial information for its description. For gene information, i.e. the gene name and gene symbol, the GEMS uses concepts from the controlled vocabulary provided by GO. The consistency for anatomical and temporal terms underlying pattern description is provided by the DAOZ, describing the developmental anatomy of the zebrafish (cf. 2.1.3) (Belmamoune and Verbeek, 2007).

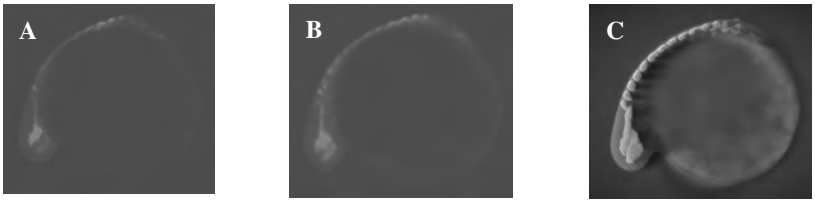

Figure 1: The zebrafish gene expression pattern of the *MyoD* gene at 24 hours post fertilization (hfp). (A) shows one slice of a 3D image, (B) shows a projection of the 3D image onto one 2D image and (C) a 3D visualization of both channels of the 3D image, i.e. the pattern in the context of the whole zebrafish embryo. The expression gene, i.e. *MyoD* is involved in development of vertebra and musculature.

In order to understand development, pattern formation as well as the analysis of patterns of gene expression, we have developed a 3D digital atlas of zebrafish development (cf. Figure 2) (http://bio-imaging.liacs.nl/liacsatlas.html). This atlas serves as a reference framework for researchers. Additionally, it is intended to serve as a model to map patterns of gene expression. In the 3D atlas, canonical developmental stages are completely described as a volumetric model in which every volume element is attributed to an anatomical structure. The atlas is built from serial sections processed to visualize standard histology. In the atlas, each 3D model is the result of 3D reconstruction of section images in which each anatomical structure is delineated by a contour; a.k.a. the graphical annotation and each anatomical domain is associated with an anatomical name representing the semantic annotation (Verbeek et al, 1999, 2000 and 2002).

Underlying the 3D digital atlas and the GEMS is the DAOZ that provides consistency for the anatomical terms as well as the temporal "staging" information of the embryo. Consequently, the GEMS and the 3D atlas are addressed with the same unique terms. Importantly, for the understanding of the pattern, specimen preparation, probe information and imaging conditions are also stored as separate annotations to the 3D image. We consider this information fundamental for comparison in the retrieval phase.

Furthermore the system assigns several system metadata to further enhance organization and management of the data.

The microarray technology is based on the analysis of expression of thousands of genes simultaneously but can not provide accurate spatial information on where a gene is expressed. Whole mount (fluorescent) *in situ* hybridization experiments enable visualization and localization of a limited number of gene expression patterns at a time. This is due to, amongst other things, limitations on the number of CLSM channels that can be used simultaneously. Combinatorial relationships, as embedded in ontologies, support retrieval of information and allow interpreting several "*in situ*" experiments. The annotation of GEMS data on the basis of structured ontologies is therefore quintessential for both retrieval and mining. So, providing access to collections of well annotated gene expression patterns is a very powerful feature of the GEMS and it contributes a solution to throughput issues that exist in (3D) whole mount *in situ* hybridization. Within GEMS we are able to cluster several co-expressed/localized patterns and, in this manner, it can be possible to reveal relations between genes. Moreover, the uniform annotation of the gene expression data with ontology terms allows cross-reference with other resources thus facilitating cross-model system comparative studies and analysis of patterns gene expression.

The specific aim of a spatio-temporal information management system for gene expression is that it refers to all relevant information within a model system, i.e. zebrafish, and directly links to all relevant other information systems, i.e. biomolecular databases. In that manner, a semantic model is built while adding information to the system. We have taken this into account in our design and implementation of the GEMS; it establishes an integrative spatio-temporal information system by making expression patterns available to the scientific community as well as interoperable with other bioinformatics resources.

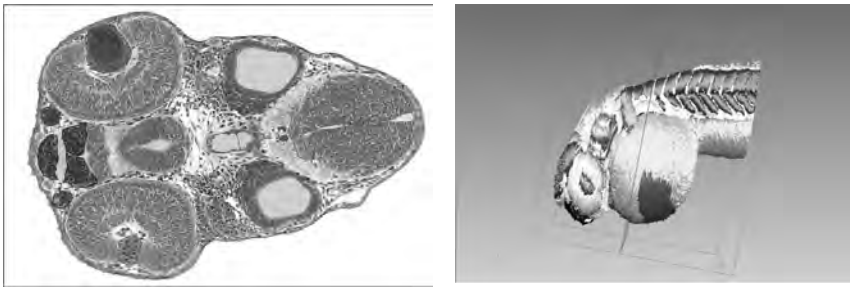

Figure 2: A 48 hpf zebrafish embryo from the 3D atlas. (Left) 2D section where each anatomical structure is annotated: semantically with an anatomical term from the DAOZ and graphically by a contour. (Right) A 3D model: resulting from 3D reconstruction from serial sections.

4.2 Material and Methods

In this section, we discuss the pattern annotation, the metadata and the system architecture in two separate subsections.

4.2.1 Pattern annotation

One of the important advantages of storing data in a database is the aptitude to query the data and compare results in an accurate manner. The accuracy can however, only be achieved if the stored data itself is accurately controlled. Comparison of biological data is often hampered by the lack of nomenclature standards. We used ontologies (Bei et al, 2006 and Gkoutos et al, 2005) to provide a standardized set of *in situ's* (zebraFISH) images prepared with the same experimental protocol and described with common nomenclature. We used DAOZ to provide a common vocabulary of terms to describe the different anatomical features of the zebrafish embryo and the different stages of embryonic development. Furthermore, we used GO to describe the expressed genes in the images.

The standard workflow for annotation of a gene expression pattern is to establish the experimental conditions, then the official gene name and subsequently finding when and where the gene is expressed. In terms of workflow, this is in a way working backward; as we know the developmental stage, the material as well as the probe in the experiment. The experimental parameters are the noteworthy information to have available. The aforementioned workflow will be further elaborated in the next three sections.

Protocol and Imaging conditions

In the context of organizing data with controlled annotations, we adhere as much as possible to standardize FISH and imaging protocols as a basis for accurate and complete pattern annotation. 3D patterns of gene expression are generated through the zebraFISH. Specimen and image preparation as well as the imaging settings represent crucial information to understand and compare the images containing patterns of gene expression.

The "*in situ*" hybridization protocols as well as the imaging settings parameters are part of the submitted data. By including the experimental data as part of the submission, we are able to keep track of small adoptions to the protocol. The submission follows the standard protocol for which XML (http://www.w3.org/) templates were created.

Gene Ontology

Patterns of gene expression in the images are annotated with terms from one of the most comprehensive ontologies within the bioinformatics community: the Gene Ontology (GO; Gene Ontology Consortium, 2006). The GO project is a collaborative effort to construct and use ontologies to facilitate the biologically meaningful annotation of genes and their products in a wide range of organisms. The GO provides an organized vocabulary, or ontology, for the description of attributes of genes and gene products, in

three key domains that are shared by all organisms, namely molecular function, biological process and cellular component.

Using the GO vocabulary contributes to data integration and in particular of the 3D patterns of gene expression with other repositories. It helps in the clarification of relationships among genes in zebrafish and between zebrafish and other model organisms. The GO is regularly updated which allows the 3D patterns annotation with up-to-date GO terminology. To extend our data integration we are investigating the possibility to use, as a supplement the species-specific annotation sets provided by GO annotation (GOA) maintained separately from the GO at EBI (http://www.ebi.ac.uk/GOA/). The GOA set will enable GO terms in GEMS to be related to other species-specific resources.

Developmental Anatomy Ontology of Zebrafish

The Zebrafish Information Network, i.e. ZFIN provides an approved and standard anatomical vocabulary of zebrafish. The DAOZ restructured the ZFIN vocabulary by introducing new concepts and relationships. The DAOZ is a task-oriented ontology for annotation. Its concepts and relationships are organized in a database system, i.e. DAOZ database, used to describe the spatial, temporal and functional characteristics of the 3D images of both the GEMS and the 3D atlas. DAOZ concepts and relationships are also used for retrieval. Therefore, images sharing similar pattern characteristics such as spatial, temporal and/or functional information could be grouped into large units which are very important for further comparison and analysis.

A substantial part of the anatomical terms are similar for each species; therefore using DAOZ terms annotation enables searching one database and also switching transparently from one species to another. In some cases, a mapping of anatomical concepts is necessary and tools to realize such mappings have been implemented (Luger et al, 2004).

A mapping can be necessary due to differences in developmental timing or the due to homologous relationships (e.g. gill vs. lung). Patterns from GEMS could, in principle be integrated with other developmental databases for comparison. In this manner the strength of the zebrafish as a model system could be better explored.

4.2.2 System administration

The GEMS manages images storage and retrieval based on administration metadata. These metadata include settings about who are allowed to submit, view and/or manipulate the data. To date, the system distinguishes three groups of users, i.e. guest users, group members and system administrators. The group members are organized per laboratory. Users having access right are allowed to submit data to the system. They are also allowed to view all data stored in the GEMS. Unregistered users are limited to view images based on permissions assigned to the images.

4.3 Implementation

In this paragraph we address the design and system architecture. The GEMS (Belmamoune and Verbeek, 2006) is implemented using the Model –View-Controller (MVC) as a design model (cf. Figure 3) and Java as programming language The MVC design model implies that the GEMS consists of three components: a central *Model*, *Views* that represent the model to the user and *Controllers* of the model.

The *Model* is where the logic of the application resides—including the data Model and any proprietary processing that must be applied to this data.
The *View* is the application as the user observes it, i.e. the layout or Graphical User Interface (GUI) in which the user can enter data into the program, get feedback and view/explore results. The GUI of GEMS is an HTML/XML/XSL based visual display interacting with the server through servlets and java server pages

(http://java.sun.com/products/jsp/).
The *Controller* is responsible for responding to user actions. In our case, a user action is a page request. The controller determines what request is being made and responds appropriately by triggering the model to manipulate the data appropriately and passing the model into the view.

Through the application model, the databases are accessed. The GEMS uses three databases. The main application database contains information about users and system administrators. This information is needed to manage access to the system. Additionally, this database includes tables with descriptive metadata of the images, image URL's as well as experimental protocols. The other two databases contain the annotation resources of the GEMS, i.e. DAOZ and GO.

We are using PostgreSQL (http://www.postgresql.org/) as the Relational DataBase Management System (RDBMS) for the GEMS while MySQL (http://www.mysql.com/) servers are used as the RDBMS for the DAOZ and GO databases. We opted for PostgreSQL which offers a high transactional performance, to allow complex queries to be performed against the GEMS database. The GO database is straightforwardly built from MySQL dumps as downloaded from the GO website (http://geneontology.org). This was our main motivation to use MySQL as the RDBMS for both annotation databases. Both GO and DAOZ databases are periodically updated.

The experimental protocol is submitted as part of pattern annotation. Its submission follows the standard '*in situ*' protocol. This protocol is always subject to change as improvements are applied. Therefore, a flexible storage format is needed. To this end, we used XML format so that a markup document could be designed and altered according to our specific needs. XSL (http://www.w3.org/) is used so that a web-browser can be used to present the XML data. For the ZebraFISH protocol, XML templates were created. The XML data is presented as a fill-in form where the user only has to indicate if deviations

were made from the standard. From these fill-in forms, the system generates XML files that represent the lab-notebook on the experimental protocol.

The 3D CLSM images are of high resolution (>= 1024^2) and can be up to 300 MB. We therefore have chosen the file system of the server above the database as a central repository for these 3D data. The image URL's are subsequently saved in the database. Consequently, the GEMS database would not grow as dramatically as it would if we stored the images in a BLOB field. The experimental protocol, i.e. XML files are also stored in the file system. From the CLSM (e.g. Leica) the imaging conditions are obtained, i.e. acquisition properties in an XML format. Through the GEMS submission interface, these data are uploaded with the images and is also stored in the file system. The GEMS database is designed to contain storage tables for 3D images descriptive metadata. When a submission is accomplished, the system automatically notifies the system administrators by sending an email containing information about both the submitted data (e.g. the assigned IDs) and the submitter (e.g. Name and group). This notification enables data tracking and review which is essential for quality control.

The GEMS as a repository, is suitable for discovery of relations between genes. To that end, we implemented pattern recognition agents in java. These applications are connected to the database and after each submission the results are updated autonomously (cf. chapter 6). This shows that the GEMS is a true dynamic repository.

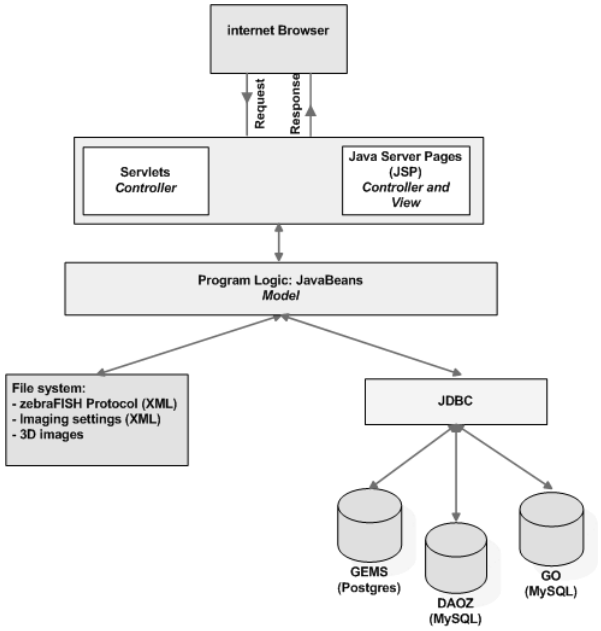

Figure 3: The Model-View –Controller architecture of the GEMS.

4.4 Results

DAOZ as well as GO are based on the biological vocabularies and establish precise defined relationships between concepts. 3D patterns of gene expression — annotated using ontology-based concepts — inherit all characteristics and relationships that these concepts possess. This situation is exploited to cluster several "*in situ*" experiments and obtain information on co-localization/expression of genes which is necessary for data interpretation. Patterns could be clustered using tissue information, i.e. spatial, functional information and/or stages of development.

We implemented methods to search the gene-expression data. These methods enable construction of SQL queries by interactive selection of concepts (cf. Figure 4). The first

method (https://bio-imaging.liacs.nl/gems/jsp/SearchImages.jsp) can be utilized to search for the expression patterns of an individual gene of interest. (e.g. “Provide me with expression information of gene X (e.g. *fgf8*)?”) (cf. Figure 4A). The second method (https://bio-imaging.liacs.nl/gems/jsp/CombinatorialSearch.jsp) is used to cluster sets of patterns using information such as: functional system, spatial location and/or stage of development (e.g. “What patterns are co-expressed in head, belong to the central nervous system and present at *prim 5*?”) (cf. Figure 4B).

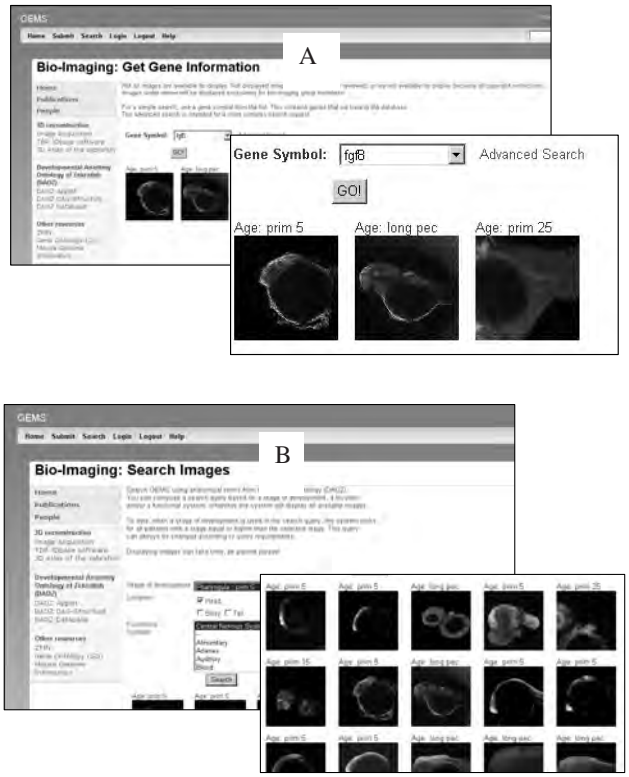

Figure 4: Search results from: (A) Search option using gene symbol. (B) Search option using concepts from DAOZ.

Sets of genes and their corresponding expression patterns can also be grouped on the basis of combinatorial search actions based on patterns spatial locations, functionality and stages of embryonic development. The system processes the constructed queries and returns image thumbnails. Each thumbnail image represents an active sub-query to generate an interactive page. This page will contain more details about the expression pattern as well as active links to related information in external databases (cf. Figure 5). On the level of database administration there exist always the possibilities of free form SQL queries using the PostgreSQL transactional features.

The 3D volume images of zebrafish embryos in both the GEMS and the 3D atlas are annotated with the same spatial and temporal information, i.e. DAOZ concepts which enable both systems integration. The integration of biological process data from GEMS and phenotypic data in the 3D atlas is usually employed to study and analyze development. To illustrate this situation we designed and implemented the 3D Visual Query system (3D-VisQus) (Belmamoune et al 2006). This system allows mapping patterns onto the atlas. It represents an advanced form of querying the GEMS through atlas 3D models.

The 3D Atlas Browser (http://bio-imaging.liacs.nl/liacsatlas.html) with a number of visualization methods enables 2D/3D atlas data to be explored. The 3D-VisQus is an extension to that atlas browser. It allows query generation based on the visual understanding of the data. Queries can be composed based on semantic and graphical annotations of the visualized atlas images. This allows users to search the GEMS for appropriate 3D patterns of gene expression (cf. Figure 5). A 2D view of one particular stage represents a user interface portal with two major features. On one hand, it is used as a browser for section images to get more insight in the 3D models anatomy. On the other hand, a selected anatomical domain in a section image is used to generate a query such as: “What patterns have gene expression in spatial location X at time Y?” which is then

executed against the GEMS database. The 3D-VisQus acts therefore as an easy portal from the atlas to the GEMS.

On the basis of its annotation, the GEMS serves as a directory to gene and spatio-temporal specific information for other databases (cf. Figure 5). The official gene symbol, as extracted from GO database during pattern annotation, facilitates direct links to several databases. In our case we provide a direct link to *Entrez Gene* database. *Entrez Gene*, developed by the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov), is a repository containing gene specific information. Each *Entrez Gene* record represents a single gene from a given organism, and provides a wide range of information such as nomenclature, chromosomal localization, gene products, markers, phenotypes, molecular interaction and many more. For the GEMS, each 3D image of an expression pattern is directly linked to a summary page within NCBI's *Entrez Gene*. The summary page allows gene-specific links to many other resources within and outside the NCBI's *Entrez* system. This integration can be exploited to process the information stored in these databases and retrieve relations in the data through machine learning mechanisms.

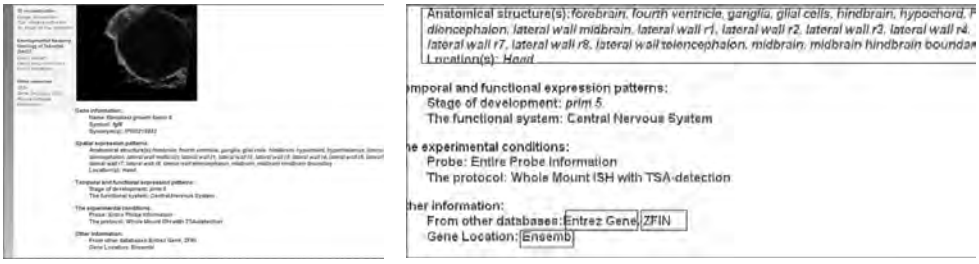

Figure 5: (Left) The detailed view of a datasheet of a pattern of gene expression provides active links (right: indicated with red boxes) for related information in ZFIN, Entrey Gene and Ensembl.

DAOZ spatial and temporal concepts allow gene expression patterns in GEMS to be directly linked to gene expression data stored in ZFIN which maintains references of

zebrafish research information and links to other model organism databases. The development of our system is complementary to ZFIN and fully interoperable with it. Finally, to map the expressed gene to a more detailed gene report and genomic location, a direct link is established to Ensembl database (http://www.ensembl.org/).

GEMS is a system in development. To date, it contains more than one hundred gene expression patterns and submission of data to the system will continue; GEMS is made available to several research groups for data submission. For information retrieval, unregistered users do not need access authentication or specific software to address the GEMS.

4.5 Conclusion and Discussion

Identifying the temporal, spatial and functional expression patterns of genes during development is a critical initial step to understand genetic networks that underpin embryogenesis processes. In this chapter we presented our approach to organize 3D gene expression data within an integrative platform, i.e. the GEMS and showed examples on how it is made accessible for researchers.

The microarray technology based on simultaneous study of co-expression of (tens of) thousands of genes; however, this technique lacks sufficient and precise spatial information of the expression domain. FISH provides the spatial visualization and localization of gene expression at a given time point, its throughput, however with respect to the number of genes is low. With the GEMS we were able to bridge the gap of time and space issues of gene expression patterns. Patterns of gene expression are annotated with time and space binding concepts allowing more experiments to be clustered as co-localized/expressed microscope experiments.

The GEMS includes well-structured and approved domain ontologies, i.e. the DAOZ and GO. The DAOZ is used to annotate 3D spatio-temporal data in the GEMS and the 3D digital atlas of zebrafish. Data from GEMS and the 3D atlas can therefore be addressed with the same unique concepts. It becomes possible to move in a seamless way from one system to the other and this situation has been illustrated with the design and development of the 3D-VisQus. This system demonstrates clearly the interoperability between patterns of gene expression in GEMS and 3D models in the 3D atlas database.

GEMS interoperability is of course not limited to the atlas but it is easily extended to other resources. The DAOZ includes similar anatomical nomenclature as the Zebrafish Anatomy Ontology which facilitates the integration of GEMS and ZFIN data. The integration of the GEMS with other resources is extended with GO terminology. GO terms enable integration of GEMS objects with a range of bioinformatics resources such as NCBI and Ensembl. Cross-references with these resources imply integration with a wealth of bioinformatics databases leading to an increase of scientific benefit of our data. To enhance data exchange with a broad range of resources, research has been initiated to extend our patterns annotation with the species-specific annotation sets provided by GOA and to use a Distributed Annotation System (DAS; http://biodas.org/) server to enable a straightforward integration of GEMS gene expression information into Ensembl.

Our study shows an example on how to design and implement an integrative information system for 3D spatio-temporal gene expression patterns and go beyond just storing and retrieving data, i.e. clustering and integrating information. Co-localized/expressed genes could be clustered given their spatial and temporal information. This approach will pave the way for finding new relations between genes. To that end, machine learning techniques can be applied to GEMS data. We have started experiments with existing algorithms to derive association rules between frequent item sets (Lee et al, 2003). Given that an association exists, one might easily infer that the genes involved participate in some type of network/pathway. Identification of such association rules is our initial step

toward analysis of genetic networks. The experiments and evaluation of the algorithms in this context are part of our future work.

Several search query interfaces have been made available. The actual search results are displayed as image generated from visualizations as projections. This operation is applied to show whole mount multi-channel content in one single color image. 3D images from the CLSM slices can be further processed with TDR-3Dbase software (Verbeek et al, 2000 and 2002). This software produces 3D models and visualizes the spatial characters of gene expression pattern. These 3D models can be viewed with our dedicated 3D browsers (http://bio-imaging.liacs.nl/liacsatlas.html) that we will embed in the GEMS for additional 3D gene expression pattern visualization

The GEMS has been successfully applied for zebraFISH 3D data storage and retrieval. The GEMS has been designed to be flexible for extension. More protocol templates, such as GFP-like as well as additional annotation resources must be embedded in the system to allow its usage in a broader context.

CHAPTER 5

3D-VISQUS: A 3D VISUAL QUERY SYSTEM INTEGRATING SEMANTIC AND GEOMETRIC MODELS

Based on:

M. Belmamoune, E. Lindoorn and F. J. Verbeek.
3D-VisQuS: A 3D Visual Query System integrating semantic and geometric models.
InSCit2006 (Ed. Vicente P. Guerrero-Bote), Volume II "Current Research in Information Sciences and Technologies. Multidisciplinary approaches to global information systems", pp 401-405, 2006.

Abstract

For developmental studies, a wealth of anatomical data is being generated. This data is increasingly complex and is generally distributed across different systems. Our anatomical data is spatio-temporal having different levels of abstraction and distributed over different database systems. This kind of data impels to search for methods for its access and analysis in a usable way. Here we try to provide a solution to access this complex data by combining easy ways for data perception with information retrieval techniques. We introduce a visual query system for our zebrafish spatio-temporal resources. With this system, i.e. 3D Visual Query System (3D-VisQus) we strive to afford end-users with a simple and intuitive interface to visualize, interact and query complex anatomical data. The 3D-VisQus is a portal to simplify users' access to complex anatomical data and to facilitate data analysis and understanding.

5.1 Introduction

For developmental studies, a vast amount of experimental 3D spatio-temporal data is being produced and managed in different database systems. We expect that these data production will continue growing and from a scientific perspective these wealth of data creates unordinary opportunities for developmental studies. The challenge that we are facing is how to quickly and effectively turn this data into valuable knowledge for a wide range of users. In this chapter we describe our solution by developing communication processes that will bridge conceptual differences across our database systems.

We developed a spatio-temporal framework to study embryonic development of zebrafish model system. To this framework we introduced an additional component, i.e. the 3D Visual Query System (3D-VisQus). It is a prototype of a query system that provides a portal interface to the framework's main components, i.e. the 3D digital atlas of zebrafish (cf. chapter 3) and *in situ* gene expression patterns organized in the Gene Expression Management System (GEMS; cf. chapter 4). The 3D-VisQus uses ontological concepts of the Developmental Anatomy Ontology of Zebrafish (DAOZ; cf. chapter 2) to link phenotypic data of the 3D Atlas to bimolecular data in the GEMS.

The 3D-VisQus is an experimental approach that we developed to overcome problems related to accessing and searching 3D spatio-temporal anatomical data. Not all users have a direct knowledge of the anatomy, yet we want them to make use of our systems. Therefore, we investigated the possibility of designing a system to provide users with an environment for data visualization, integration and retrieval. Some of its required properties are: visual representation of 3D spatio-temporal models, different level-of-detail definitions when browsing in a graphical model and the possibility of representing a textual query in a visual way. We designed a prototype of such system, i.e. the 3D-VisQus to access and query our 3D spatio-temporal anatomical information. 3D-VisQus will be considered as a mental model that translates users' thoughts processes for how

anatomical structures are related to each other into real 3D models. Moreover, it will assists users to formulate readily their search queries using visualized graphical data while underlying systems and the query language are transparent to users.

The concept of using visual perception to query data has already been introduced, e.g. geographical information systems (GIS). In these systems users are able to express their queries in a visual environment. Users formulate their search queries in a visual way by selecting appropriate image icons and putting them into the proper places. To facilitate information understanding the results of a search query are also presented in a graphical manner (Goncalves et al, 2000). In our framework we have different kind of data, i.e. 3D spatio-temporal models. The system should first visualize 3D models and second translates users' actions to formulate search queries.

The 3D-VisQus is based on the 3D Atlas Browser (cf. chapter 2) which was designed to visualise and explore anatomical data of zebrafish model system in 3D. Such 3D data gives users the opportunity to explore anatomical structures in the correct manner. The 3D-VisQus extended the 3D Atlas Browser to use a visualized 3D model as a query interface to search for patterns of gene expression. We are applying the concept of visual queries by using each anatomical domain in a visualized image as an abstract object to generate a query. Anatomical domains are annotated graphically and are labeled with a textual annotation. Therefore from choosing a graphical domain, the system extracts the spatial and temporal information of that domain and generates a search query clause. The query is then executed against the DAOZ and GEMS database. The 3D-VisQus prototype is implemented on top of our framework databases. The 3D-VisQus is, therefore, a portal to our database resources to which a typical retrieval environment is attached. Searching with a system such as the 3D-VisQus will make querying complex data more intuitive and users do not have to worry about query languages nor the underlying databases.

The 3D digital atlas of zebrafish (cf. chapter 2) is used as a spatial reference framework for mapping data. It is our representation of zebrafish anatomy that provides a visual standardized coordinate system for the search and analysis of expression data. 3D patterns of markers genes expression are obtained from whole mount in situ hybridization (Welten et al, 2006) and are organized in the GEMS database (cf. chapter 4). The gene expression database must describe the time and space of gene expression in a standardized way. To achieve this goal, we used DAOZ (cf. chapter 2) for common nomenclature for data annotation in GEMS and 3D atlas models. This common annotation enables mapping GEMS data onto 3D atlas of zebrafish development.

In the DAOZ anatomical structures are modeled hierarchically from functional system, body region to substructures. Each anatomical concept could have more than one hierarchical parent using more than one relationship specifying in which way these concepts are related to each other. This hierarchy provides different level of granularity that facilitates data organization, analysis and retrieval. The ontology concepts represent the pivot of the 3D-VisQus to integrate the 3D atlas and GEMS components and to compose and execute search queries against the 3D atlas and GEMS database systems using the different levels of data abstraction. Through the visual query interface of the 3D-VisQus users are able to compose spatio-temporal queries when interacting with the visualized graphical entities. The system extracts the spatio-temporal information of the region of interest and generates metadata. These represent the basic components of the queries that the system creates and executes against our databases models, e.g. “select all patterns of gene expression where the expressed patterns are located in anatomical domain X and present at time Y”. By this system a mapping between gene expression data, the 3D atlas and DAOZ is realized and moreover a novel paradigm for searching based on visual understanding of data is introduced and explored.

5.2 3D-VisQus Usability

Our anatomical data could be considered as complex since it is 3D, spatio-temporal and distributed over different database systems. The learnability concept is the key of a successful system such as the 3D-VisQus to access and search these data. The system should therefore be easy to lean with an intuitive interface. This interface should present anatomical information with the focus being to make such information easy to access in order to sufficiently engage the user. Anatomical data should be available in different views and scales. Multiple view approach is wished for users to deal with related data at different levels of abstraction. The points of interest should be visible in both global and detailed views; related aspects only visible, but not detailed, until the user chooses them. In this context of these usability requirements, the 3D-VisQus has been designed.

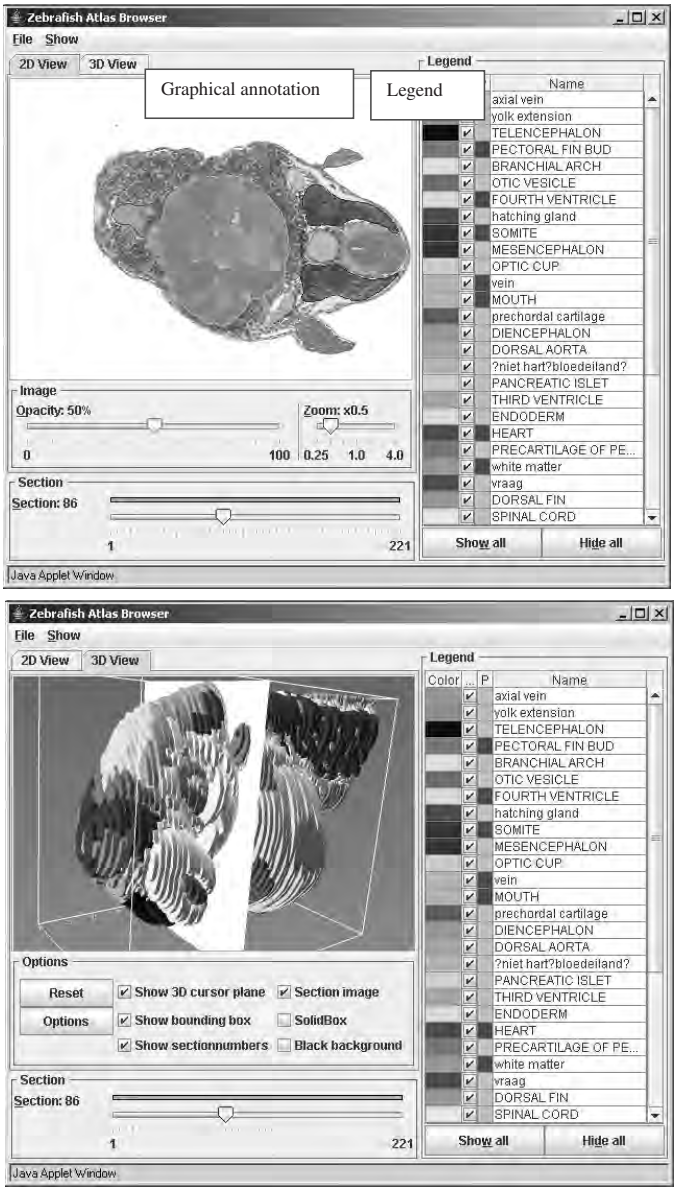

Figure 1: (Above) 3D atlas: 2D and (Below) 3D views of a zebrafish embryo at 36 hours post fertilization (hpf).

The 3D-VisQus is an extension of the 3D Atlas browser (cf. Figure 1). The extension resides in 3D-VisQus capability to compose queries based on user's perception of the data. We have explored the possibility to use section images of a 3D model as a query interface. The system uses a low level of data visualization in the form of images, i.e. the Graphical User Interface (GUI) to retrieve high level elements in other database systems.

Actually, the 2D view of a visualized 3D model represents a portal user interface with two major capabilities. On one hand it is used to explore a 3D model and the section images to get more insight of the 3D data. On the other hand each section image within the 2D view is used to generate a query, which is then executed against the underlying databases, i.e. DAOZ and GEMS databases. Within a section image, each anatomical structure is bounded by its contour representing the graphical annotation to which an anatomical name is associated that represents the semantic annotation. Each anatomical domain is associated with a specific label that facilitates distinguishing and highlighting information of interest (cf. chapter 2).

The 3D-VisQus offers two query forms. First, users are able to query into the visualized model itself. In a 2D view, visualizations are synchronized. In the legend at the left pane (cf. Figure 1), when an anatomical name is selected, the main pane is immediately updated with the first section image containing the anatomical domain related to the selected name (Verbeek et al, 2000 and 2002). Second, each anatomical domain is used to query the GEMS for patterns with expression in that domain. The query is an association between the spatial and temporal characteristics of the anatomical structure. The spatial information is provided by both the graphical and semantic annotations of the visualized anatomical domain while temporal information is obtained from the developmental stage of the visualized embryo (3D model). When an end-user wants to make a query, there are three main steps: what to query, where to query and how to query. In the following, we give an example to illustrate the visual query process. Assume an end user wants to know expression patterns within structure X at

developmental stage or time Y. First, end user selects the proper 3D model (where to query) which is then displayed with the 3D-VisQus. Second, the user looks for the structure X in the displayed section images (what to query) and interacts with (clicks on) the anatomical domain of this structure to generate a system action. At this point, the system, i.e. 3D-VisQus translates this operation to system's internal query in the form of: "select all patterns of gene expression in structure X and at time Y". This query is then executed against the underlying database systems and the patterns of gene expression are displayed.

The system is based on three spatio-temporal database management systems (DBMS) (cf. Figure 2). The main spatio-temporal database is this of the 3D digital atlas. The visualized 3D models are stored in this database while queries are executed against the other two databases: the DAOZ database and the GEMS database.

Visualization of the query result is critical to the end user. The output is given as a combination of two forms: graphical and textual. The system is able to translate textual results to a graphical output. This is given in the form of thumbnails of patterns of gene expression that are generated in real time. Each thumbnail image is a 2D projection of the original multi-channel in situ hybridization images stored in the GEMS (cf. chapter 3). Additionally, each image carries expression information of a given gene and is indicated with this gene symbol representing the textual information. To explore data, end users can move from global overview to more detailed information. This principle is applied to explore the data and is also respected in the visualized output data. The output list provides a global visual data analysis. At this stage, the system also provides details on demand. When a user selects a thumbnail image a second query is submitted to the underlying system. A detailed view about the selected pattern is then generated in real time (cf. Figure 4).

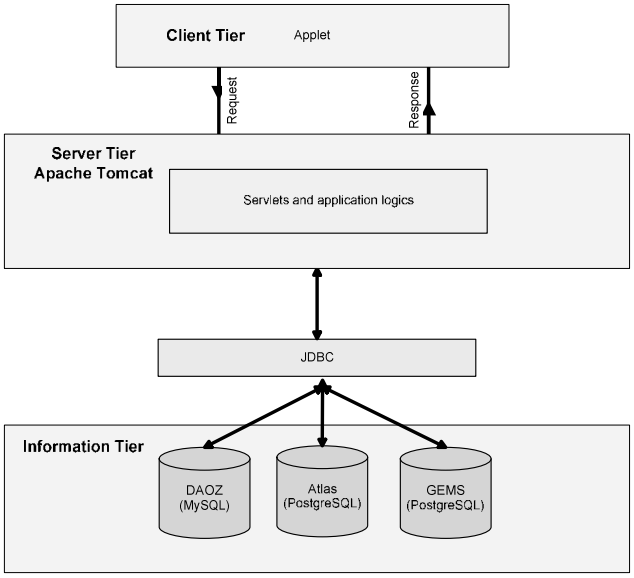

Figure 2: The 3 tier architecture of the 3D-VisQus.

In accordance to the architecture of the atlas, the DAOZ and GEMS, we also adopted a multi-tier architecture for the 3D-VisQus application (cf. Figure 2). 3D models are extracted from the atlas database and are visualized with a java applet that can be deployed on the client side. We are using the applet as the application front-end. The applet communicates with the server side through a servlet. We need to capture the user action and pass this information to the servlet. Since servlets support the HTTP interface, we communicate with the servlet over HTTP socket connections. The applet opens a connection to the specified servlet URL. Once this connection is made, then the applet can get an output stream or input stream on the servlet.

The applet translates the user's action to a request and sends this to the server tier. This tier processes client requests and sends results back to the client. The server tier formulates and executes the query based on the action type, the spatial feature of the

region of interest and the temporal information, i.e. development stage of the visualized atlas model. Both the DAOZ and the GEMS databases are addressed. The first is queried to get control identifiers of the spatio-temporal concepts of the interested region while the second uses this information to retrieve patterns of gene expression responding to these search criteria (cf. Figure 3).

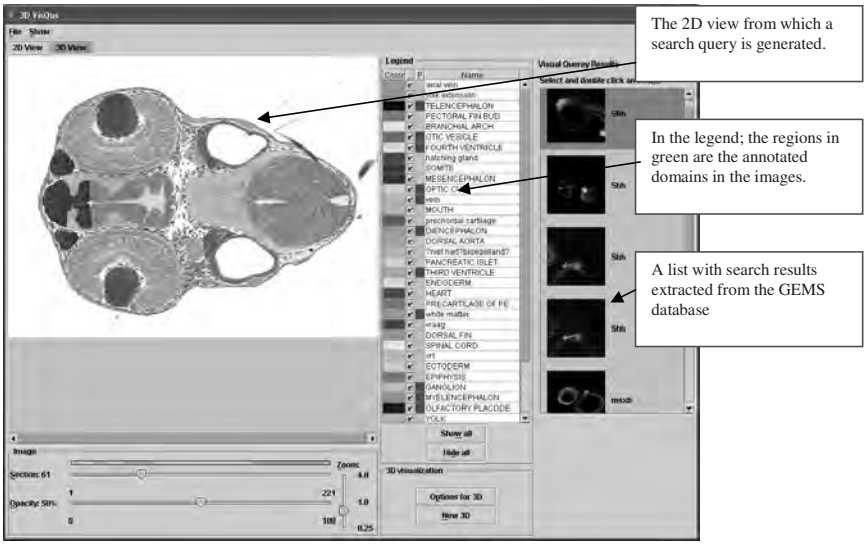

Figure 3: A search example of 3D-VisQus. By clicking on a region of interest in the visualized section image the system creates and executes a query against the GEMS and the search results are displayed. If there is a void query result, the system looks in the proximity location of the indicated anatomical domain. In this case the following query has been generated: "select all patterns with gene expression domain located in head and not older than 48 hpf".

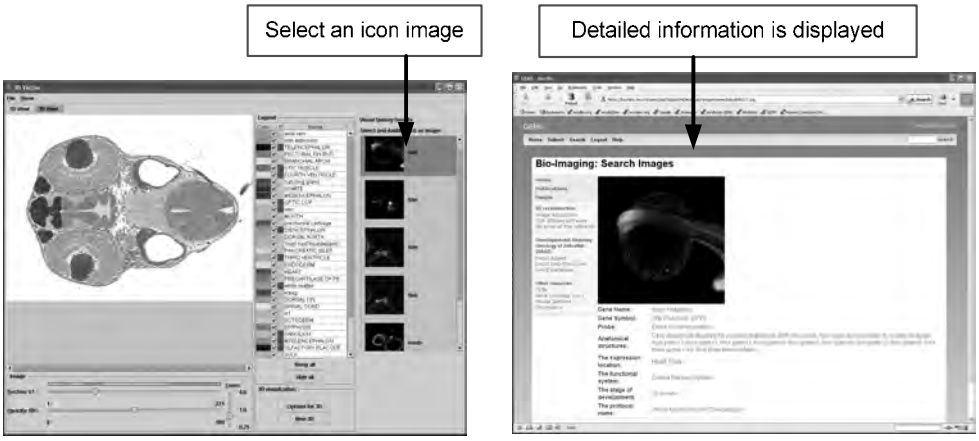

Figure 4: From the output list (left) a detailed view of a selected pattern can be generated (right).

5.3 Users Analysis and System evaluation

There are two main groups of target users comprising creators of the 3D models and biologists interested on the 3D data. The first group is of experienced biologists. They create atlas 3D models using the TDR-3Dbase (cf. chapter 2). The second group represents our main audience; they could be experienced biologists as students interested on using our spatio-temporal data. For this group, it is important to access the information as quickly and easily as possible.

To investigate how the 3D-VisQus will be accepted, we focused on measuring its capacity to meet its requirements. Therefore a set of task scenarios were developed. These represent realistic situations wherein the tester performs a list of tasks. The evaluation was carried out with a small group of target users comprising collaborator biologists and novice users. The evaluation involved a walk through the developed scenarios to capture typical user tasks in a normal work flow and to allow users to

explore offered functionalities. Based on uses feedback, we concluded that the system meets the most important requirements. Users also made suggestions for improvements and changes to the system. Such as enabling query storage for later use, and the possibility of representing a visual query in a textual way as a form of feedback to the user.

5.4 Conclusions and future work

Integration of spatio-temporal gene expression patterns with the anatomy from the 3D atlas is important for genetic and developmental studies. It is crucial for researchers to be able to access, search and combine such information for an effective understanding of the anatomical development. The 3D-VisQus is an experiment that we derived to investigate a mechanism to intuitively map genotype into phenotype data. With a system such as the 3D-VisQus we could realize a primary mapping between in situ data of different expressions patterns into the 3D digital atlas. Moreover, the 3D-VisQus offers many other advantages. It provides the possibility to explore 3D data and dynamically formulate correct and exact visual query clauses. When compared to a semantic query system, this form of data retrieval offers a great flexibility in formulating complex search queries. In our case, users are not required to have a deep knowledge of development in order to formulate a search query with exact anatomical terms. With a few maneuvers (mouse clicks), users formulate their search queries based on their visual perception and data recognition. The 3D-VisQus extracts and executes text-based SQL statements from the submitted visual query while the underlying databases models and the query language are completely transparent to the users.

The 3D-VisQus is our proposed prototype of an interface to visualize and query our complex anatomical data. The 3D-VisQus is a portal to our database resources. The system allows search based on visual understanding of the data, and as such bridges conceptual and linguistic differences between systems as well as users. Early usability

evaluation involved expert and novice users. This primary evaluation revealed that users appreciate the interface and the possibilities it offers to query spatio-temporal databases in an intuitive manner. Therefore, we are encouraged to improve this system to be used as a valuable visual query system on the top of our database systems. We need to add advanced search methods by taking full advantage of the combinatorial relationships among anatomical structures and to mirror these in the search capabilities of the system. In this initial prototype the focus was on the spatial relation: the *located_at* operator of the selected anatomical domain. As a future improvement, the system needs to generate a query based on the selected anatomical entity; in case of no results the system should come up with a new query with a higher granularity level based on the spatial characteristics of the selected graphical entity. These characteristics can easily be extracted from the DAOZ. In order to describe the spatial relationships more clearly between anatomical entities, operations like *part_of*, *near_of*, *located_at* can be used. For example, if there are no patterns associated with the anatomical term 'hindbrain' then the system should be able to generate a new query using granularity and relationship covering the whole ´brain´ region by using the *part_of* relationship. This feature is made possible because of the advanced hierarchical structure of the anatomical ontology of the zebrafish (cf. chapter 4). In future embellishments the system should enable query storage for later use, and the possibility of representing a visual query in a textual way as a form of feedback to the user. The actual output of a search query is given as 2D images, i.e. patterns of gene expression, while, the ideal output should be a spatial mapping of the expression domains onto the standard atlas model. Spatial concepts from the DAOZ are used to annotate anatomical and expression domains in the 3D atlas models and patterns of gene expression in GEMS respectively. To improve the 3D-VisQus, these concepts could be reused to provide a spatial mapping between the atlas and in situ gene expression data. The 3D-VisQus is a promising on-going research project intended to contribute to our user's satisfaction.

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 $I = \{i_1, i_2, i_3, \ldots, i_n\}$ 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 implication form:

a. $X \Rightarrow Y$ 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.

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* $\Rightarrow$ *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 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.

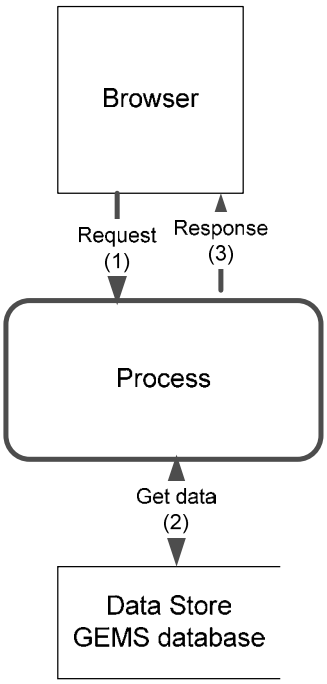

Figure 1: The process flow of the web-application to mine expression patterns.

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.

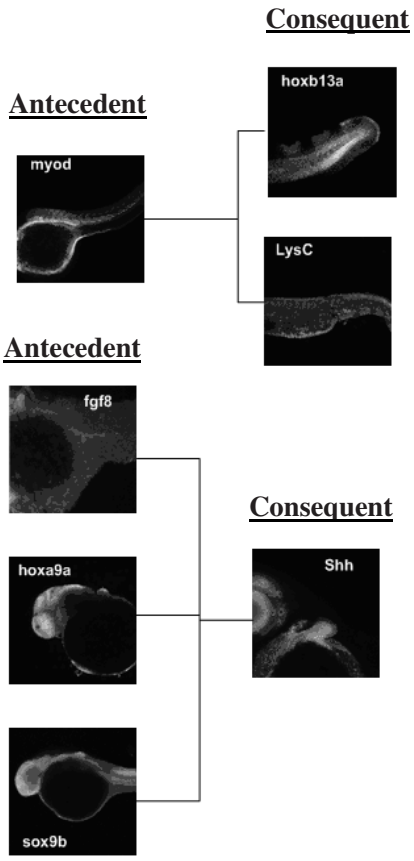

Figure 2: An example extracted from the result set of the PPM algorithm (support >=30% and confidence >= 75%) on the GEMS dataset. The first tree genes have a common expression in *tail* while the second tree contains rules with genes having a common expression in *fin* (in the body region).

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.

CHAPTER 7

CONCLUSIONS AND DISCUSSIONS

7.1 General overview

In this thesis we presented a 3D spatio-temporal framework for the zebrafish model system. This framework is intended to assist biologists in their studies of vertebrate development and is based on a number of components. The separate components were presented in the different chapters of this thesis. This chapter summarizes our conclusions and provides a discussion for each chapter separately.

7.2 The Developmental Anatomy Ontology of Zebrafish

In Chapter 2 we presented the Developmental Anatomy Ontology of Zebrafish (DAOZ). DAOZ is being developed to fill our need for an anatomy ontology to be used and adapted by computer-based applications that require anatomical information for data annotation and retrieval. DAOZ provides both a spatial and temporal structured ontology based on the anatomical vocabulary provided by ZFIN (http://zfin.org). Furthermore, we extended DAOZ with additional concepts and relations to cover the phenotypic structure of the zebrafish model organism at the most biologically relevant levels of granularity. Through the different relationships the ontology concepts are being structured as a Directed Acyclic Graph (DAG) and provided with clear semantics. DAOZ structure has been expressed in a set of axioms for a formal and consistent description. Furthermore, the complete ontology has been organized in an object-oriented database where we assigned to each ontological concept a unique identifier. Storing concepts in a database enables DAOZ to be readily navigable and understandable by curators for data annotation and by computers for data retrieval and mining. DAOZ concepts are in compliance with these known inside the zebrafish community; this assures integration and data dissemination.

An ontology can always be extended with more granularity for a larger data description and propagation; the DAOZ DAG structure enables such extensions with additional concepts and relations to improve the ontology maintenance and dissemination.

7.3 The 3D Digital Atlas of Zebrafish

In Chapter 3 we presented the 3D digital atlas of zebrafish model organism that has been restructured in an object oriented database system. This atlas contains 3D models at representative stages of zebrafish development. The 3D digital atlas serves as a coordinate framework for data comparison and analysis. Additionally, it could be a particularly useful tool for education. In our spatio-temporal framework, the 3D digital atlas is an indispensable component; to be used as a reference tool for submission of patterns of gene expression and retrieval of data it is has been restructured in a database system using DAOZ concepts as common nomenclature for its data annotation. Each 3D atlas stage model consists of a complex set of 3D volumetric anatomical structures. Anatomical structures – or domains- are annotated with unique identifier of spatio-temporal anatomical concepts from the DAOZ. Through DAOZ concepts anatomical domains of the 3D atlas models and set of images are restructured in a database system using different levels of resolution. Using the three-tier web-application, i.e. 3D ZFAtlasServer users can explore and query 3D models through the internet. This application uses the different levels of data organization to search anatomical domains in a 3D model. Queries are composed at a gross level of granularity based on general concepts to get detailed results at a finer level of abstraction. This query facility offers a readily access to this complex anatomical data for a wide range of users. Users can specify data to view according to their needs following the principle of '*you get what you want*'.

7.4 The Gene Expression Management System

Chapter 4 is dedicated to the description of another component in our framework, i.e. the Gene Expression Management System (GEMS). Molecular biology sometimes redefines anatomical borders through patterns of gene expression. The GEMS has been developed to manage, link and mine *in situ* expression patterns of marker genes during the embryonic development of zebrafish. Patterns of gene expression are 3D images; they are resulting from whole mount Fluorescent *In Situ* Hybridization (FISH) experiments, i.e. zebrafish (Welten et al, 2006) and imaging with the Confocal Laser Scanner Microscopy (CLSM) imaging. Where and when certain sets of genes are expressed regulate processes such as responses to cellular differentiation by growth factors. A number of developmental studies have already been realized using microarrays. These studies provide high throughput gene expression information at a gross level of granularity, i.e. in a mixture of cells or in whole organisms. However, the ideal situation to study development is to provide expression information at a finer level, i.e. in population of identical cells. Whole mount *in situ* hybridization is a process that enables gene expression visualization at cellular level in a whole organism. In our case, patterns of gene expression have a spatio-temporal dimension since expression information is visualized within cells (spatial information) of an intact zebrafish embryo (temporal information). The GEMS is the central repository where this spatio-temporal gene expression images are managed into the proper spatial and temporal context so that data can be adequately analyzed and thus contribution to developmental study can be usefully initiated. Most raw data from *in situ* studies are never published. With the GEMS, we facilitate online data submission and annotation of the original images from collaborating laboratories. The online submission to collect this raw data is the key for our system success.

The system allows management of several data types. The experimental protocol is part of the images submission. An experimental protocol could always play an important role for data analysis. Therefore, an image submission starts by presenting the experimental protocol as a form filled by pre-defined values that the users could always change and submit. An experimental protocol can always be modified. Therefore, we managed protocol information using XML format that offers a great flexibility for adaptation and maintenance then using a rational database. Each submitted protocol is stored as XML-laboratory notebook.

In order to support interpretation and comparison of gene expression datasets, gene expression patterns need to be linked with other resources for data mapping and comparison. The mapping of gene expression patterns onto the 3D atlas is realized using the consistent DAOZ concepts. DAOZ concepts also provide data consistency with data inside the zebrafish community. Additionally, Gene Ontology (GO) terminology is used as a vocabulary to annotate gene and gene product of GEMS expression patterns; trough the GO terminology, GEMS data integration is extended to other resources such as NCBI and Entrez gene.

The most critical issue in a database system design is information access. We provide several ways to access the information residing in GEMS. The GEMS supports a variety of search query possibilities through different Graphical User Interfaces (GUIs). In its actual release, the GEMS provides data visualization in 2D format. As it is known, it is difficult to visualize in our mind 3D data. Therefore, 3D representation of the expression patterns is needed in order to give molecular information for developmental components in a 3D context. To this end, graphical models are derived. These models in the future will be integrated into the system to offer an additional 3D visualization to end-users. The GEMS, as many other central repositories, is confronted with the task how to improve its data exchange with the rest of the community. To this end, at the moment the Distributed Annotation System (DAS) is under consideration. DAS is being used to exchange

biological annotations between data distributed among different web-sites and a system such as DAS will be very useful to improve our data exchange with other resources.

7.5 The 3D Visual Query System

Chapter 5 was focused on a query system based on visual formulation of search queries. The complexity of spatio-temporal data requires tools that support and facilitate interactive data exploration. We have built a prototype environment for interactive querying and exploration of spatio-temporal data. When formulating visual queries, users start with zebrafish 3D models. Each 3D model represents the graphical representation of the input data. Users navigate through the data to look for domains of interest. The state of the visualization environment forms a visual query. The query output is a set of patterns of gene expression from the GEMS and expression models if available. In this chapter we described the prototype of the 3D Visual Query System (3D-VisQus). As an additional component to our information framework, 3D-VisQus offers a portal and intuitive interface to GEMS database through 3D models of the atlas. This system links the two components with unique identifiers of DAOZ concepts. With the 3D-VisQus we demonstrated two key elements; i.e. (1) we showed a query method based on perception and recognition of the visualized elements facilitating access and exploration to complex anatomical data; (2) by the 3D-VisQus we demonstrated that the consistent annotation and organization of GEMS data and atlas models in well structured databases allows their integration and mapping.

7.6 The GEMS: a mining tool for spatio-temporal patterns

In Chapter 6 we explored data mining strategies and we focused on mining association rules between sets of gene expression patterns in the GEMS database. Furthermore, we focused on user interaction with the rules resulting from the classification in the data. The GEMS, i.e. spatio-temporal framework for patterns of gene expression has been extended

by introducing an additional functionality to construct and mine association rules. Within the graphical user interface of GEMS, researchers are able to run operations to mine association rules between gene expression data. Genes have different temporal expression profiles. To take their temporal characteristics into consideration, we adopted the Progressive Partition Miner (PPM) algorithm to mine association rules over annotated images in the GEMS. Furthermore, we developed java agents to execute the algorithm autonomously with each submission. Through the GEMS user interface users are able to send requests to the system to run the mining algorithm and to generate rules on the fly. The same GEMS framework is used to put the discovered rules by intuition into context with existing biological knowledge. The spatio-temporal images and descriptive annotations of the generated rules represent a first attempt for users to start data analysis. Furthermore, the GEMS framework enhances each expression pattern analysis by providing links to external resources (cf. chapter 4).

We started with the PPM algorithm to introduce and explore mining aspect within the GEMS framework. In our future work, we intend to use other promising approaches such as Frequent Episode Mining in Developmental Analysis (FEDA). This algorithm is more tailored to our developmental data as it is based on analyzing sequences of developmental characters to discover episodes and these are used to determine differences between developmental sequences (Bathoorn et al, 2007).

With this additional functionality to mine for association rules we showed that the GEMS offers a platform for linking and mining 3D spatio-temporal patterns of gene expression of zebrafish model system. We mined for association rules over annotated images. Moreover, mining for associations over features in the images is under considered (Jano et al, 2007).

7.7 General conclusions

In this thesis we presented a spatio-temporal framework to facilitate studies in developmental genetics and biology. This framework is composed of three main components, i.e. the 3D atlas of zebrafish that represents the reference framework for zebrafish data submission and retrieval, the GEMS which is the central repository for 3D spatio-temporal patterns of gene expression for data storage, retrieval and mining and the DAOZ corresponding to the standard semantic framework for data annotation in both the 3D atlas and the GEMS. Through dedicated three-tier web-applications each of these components can be accessed separately for an easy data exploration and analysis. An additional component, i.e. the 3D visual query system has been developed to serve as a portal interface to access our spatio-temporal data as a whole. It facilitates user's tasks to access GEMS data through 3D atlas models by using DAOZ entities. The study presented here is in its proof-of-principle stage that will be followed by the widespread-adoption stage to cover more model systems.

The actual developmental framework should be extended with additional components such as microarray elements. A wealth of information can be obtained from microarrays to understand genetic network through development. Microarray for genes expression profiling when combined with spatial and temporal patterns of gene expression may highlight the processes that take place during embryonic development. Comprehensive microarrays covering large numbers of the predicted expressed transcripts for zebrafish are available. Whole mount *in situ* data stored in the GEMS is typically used to provide gene function information within a biological process. Microarrays are necessary for pathways analysis and understanding the gene expression network within which a particular gene operates. Therefore, with the addition of microarrays, that all of the tools for complex dissections of both cellular and genetic pathways are available to developmental biologists. Given this data availability of arrays and our computing

infrastructure, we should make increasing use of this wealth of data to improve our framework for computational assessments. To conclude, in this thesis we presented our platform for linking and mining spatio-temporal databases of gene expression for zebrafish developmental model organism. This platform is dedicated to study zebrafish development. However, it has been designed in such a way to be scalable to cover other model organisms for an improved environment for developmental studies. Moreover, through this platform data linking and exchange with other model organisms should be facilitated and be more easy to realize. This spatio-temporal framework is an on-going research that opens up cross platform validation and search to improve developmental studies.

References

DAG-Edit: http://amigo.geneontology.org/dev/java/dagedit/docs/index.html

Distributed Annotation System (DAS): http://biodas.org/

Developmental Anatomy Ontology of Zebrafish (DAOZ):

ENSEMBL: http://www.ensembl.org/

GEMS Search:

- Using anatomy concepts: https://bio-imaging.liacs.nl/gems/jsp/CombinatorialSearch.jsp
- Using gene information: https://bio-imaging.liacs.nl/gems/jsp/SearchImages.jsp

Java Server Pages (JSP): http://java.sun.com/products/jsp/

MySQL: http://www.mysql.com/

NCBI: http://www.ncbi.nlm.nih.gov

Open Biology Ontologies site, http://obo.sourceforge.net/

Portable network graphics (png): http://www.w3.org/Graphics/PNG/

PostgreSQL: http://www.postgresql.org/

The 3D atlas of zebrafish: http://bio-imaging.liacs.nl/liacsatlas.html

The Gene Expression Management System (GEMS): http://bio-imaging.liacs.nl/gems/

The Gene Ontology http://geneontology.org

Zebrafish Information Network (ZFIN): http://zfin.org

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

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

Baldock, R. and Burger, A. Anatomical ontology: names and places in biology. Genome Biology, 6:108, 2005.

Bard, J. B. L., Rhee, S. Y. and Ashburner, M. An ontology for cell types. Genome Biology, 6:R21, 2005.

Basset, D. E., Eisen, M.B., Boguski, M.S. Gene expression informatics – it's all in your mine, Nature Genetics supplement, Vol. 21, Jan. 1999, 51-55.

Bei, Y., Belmamoune, M. and Verbeek, F. J. "Ontology and image semantics in multimodal imaging: submission and retrieval", Proc. of SPIE Internet Imaging VII, Vol. 6061, 60610C1 C12, 2006.

Bathoorn, R. and Siebes, A.J.P.M. 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.

Bathoorn, R., Welten, M.C.M., Siebes, A.J.P.M., Richardson, M.K. and Verbeek, F.J. Limb - fin heterochrony: a case study analysis of molecular and morphological characters using frequent episode mining. (Submitted for publication)

Belmamoune, M. and Verbeek, F. J. 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.

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

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

Belmamoune, M. and Verbeek, F.J. 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.

Belmamoune, M., Lindoorn, E. and Verbeek, F. J. 3D-VisQuS: A 3D Visual Query System integrating semantic and geometric models. In: InSCit2006 (Ed. Vicente P. Guerrero-Bote), Volume II "Current Research in Information Sciences and Technologies. Multidisciplinary approaches to global information systems", pp 401-405, 2006.

Boissonnat, J.D. Geometric structures for three-dimensional shape representation. ACM Transactions on Graphics, 3(4):266286, October 1984.

Brune, R.M., Bard, J.B., Dubreuil, C., Guest, E., Hill, W., Kaufman, M., Stark, M., Davidson, D. and Baldock, R.A. A three-dimensional model of the mouse at embryonic day 9. Dev Biol 1999, 216(2):457-468.

Camon, E., Margane, M., Barrel, D., Lee, V., Dimmer, E., Malsen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. D262±D266 Nucleic Acids Research, Vol. 32, Database issue DOI:10.1093/nar/gkh021, 2004.

Christiansen, J.H., Yang, Y., Venkataraman, S., Richardson, L., Stevenson, P., Burton, N., Baldock, R.A. and Davidson, D.R. EMAGE: a spatial database of gene expression patterns during mouse embryo development. Nucleic Acids Res. 2006; 34:D637–D641.

Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. D322–D326 Nucleic Acids Research, 2006, Vol. 34, Database issue.

Gkoutos, G.V., Green, E.C.J., Mallon, A., Hancock, J.M., Davidson, D. Using Ontologies to describe mouse phenotypes. Genome Biology, 6:R8, 2005.

Grumbling, G., Strelets, V. FlyBase: anatomical data, images and queries. Nucleic Acids Res. 2006;34:D484–D488.

Haendel, M. A., Neuhaus, F., Osumi-Sutherland, D. S., Mabee, P. M., Mejino, J. L. V., Mungall, C. J. and Smith, B. A preprint of the chapter 'Modelling Principles and Methodologies – Spatial Representation and Reasoning' in Albert Burger, Duncan Davidson and Richard Baldock (Editors): Anatomy Ontologies for Bioinformatics: Principles and Practice.

Henrich, T., Ramialison, M., Wittbrodt, B., Assouline, B., Bourrat, F., Berger, A., Himmelbauer, H., Sasaki, T., Shimizu, N., Shimizu, N., Westerfield, M., Kondoh, H. and Wittbrodt, J. MEPD: a resource for medaka gene expression patterns. Bioinformatics. 2005;21:3195–3197.

Isogai, S., Horiguchi, M. and Weinstein, B. M. The Vascular Anatomy of the Developing Zebrafish: An Atlas of Embryonic and Early Larval Development. Developmental Biology 230, 278–301 (2001) doi:10.1006/dbio.2000.9995

Kabli, S., Alia, A., Spaink, H.P., Verbeek, F.J., de Groot, H.J.M. Magnetic Resonance Microscopy of the Adult Zebrafish. Zebrafish (2006), Vol 3., #4, pp 431 - 439

Kimmel, C. B., Kimmel, S. R., Ullmann, B., Schilling, T. F. and Ballard, W. W. Stages of embryonic development of the zebrafish. Dev Dyn 203, 3, 253-310, 1995.

Lee, C., Chen, M. and Lin, C. 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.

Lenzerini, M. "Data Integration: A Theoretical Perspective". PODS 2002: 233-246

Luger, S., Aitken, J.S., Webber, B.L. Cross-species Mapping between Anatomical Ontologies Based on Lexico-syntactic Properties. ISMB-2004. Poster C-40. S.F. Gilbert. (2006) Developmental Biology, Eighth Edition

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

Meuleman, W., Welten, M. C., Verbeek, F. J. 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.

Patrick, J. Metonymic and Holonymic Roles and Emergent Properties in the SNOMED CT Ontology. Advances in Ontologies, M. Orgun & T, Meyer (Editors). Proc of the Australasian Ontology Workshop (AOW 2006), Tasmania. 2006.

Richardson, M.K. et al. with Belmamoune, M., Bertens, L.F.M., Verbeek, F.J. (2009) Zebrafish development and regeneration: new tools for biomedical research. Int. J. Dev. Biol. (2009) 53: 835-850.

Ringwald, M., Baldock, R., Bard, J.B., Kaufman, M., Eppig, J.T., Richardson, J.E., Nadeau, J.H. and Davidson, D. A Database for Mouse Development. Science vol. 265, 30 September 1994.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. L. and Rosse, C. Relations in biomedical ontologies. Genome Biology, 6:R46, 2005.

Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D. G., Mani, P., Ramachandran, S., Schaper, K., Segerdell, E., Song, P., Sprunger, B., Taylor, S., van Slyke, C. E., and Westerfield, M. The Zebrafish Information Network: the zebrafish model organism database. Nucleic Acids Research, Vol. 34, Database issue D581 D585, 2006.

Verbeek, F. J. and Boon, P. J. 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.

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

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

Verbeek, F.J. 3D reconstruction from serial sections, applications and limitations. Microscopy and Analysis 96-11, 33-35 1996.

Verbeek, F.J. and Boon, P.J. High resolution 3D-reconstruction from serial sections; microscope instrumentation, software design and its implementations. Proceedings of SPIE 4621, 65-76, 2002.

Verbeek, F.J. and Huijsmans, D.P. 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.

Verbeek, F.J. 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

Verbeek, F.J., Huijsmans, D.P., Baeten, R.W.A.M., Schoutsen, C.M., and Lamers, W.H. Design and implementation of a program for 3D-reconstruction from serial sections; a data driven approach. Microscopy Research and Technique 30, 496 512, 1995.

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

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

Weninger, W.J. and Mohun, T. Phenotyping transgenic embryos: a rapid 3-D screening method based on episcopic fluorescence image capturing. Nat Genet 2002, 30(1):59-65.

SAMENVATTING

Bio-informatica kan omschreven worden als het toepassen van algoritmen om meerwaarde te verkrijgen uit data afkomstig van biomedisch en/of biologisch onderzoek. In bio-informatica wordt onderzoek gedaan met grote gegevens verzamelingen die afkomstig zijn uit biomedisch en/of biologisch experimenten. Het doel van dit onderzoek is komen tot nieuwe inzichten vanuit de gegevens verzameling. Deze inzichten komen tot stand door de goede organisatie van de data, het linken naar en integreren met complementaire gegevens verzamelingen en ontwikkelen en toepassen van analytische methodieken. Als bio-informatica groep onderzoeken wij het inrichten en ontwikkelen van een 3D spatio-temporele data omgeving voor het zebravis model organisme. In deze omgeving wordt 3D anatomische data georganiseerd om de ontwikkelingsstudies te ondersteunen. In dit proefschrift ligt de nadruk op organisatie van 3D anatomische structuren en 3D patronen van gen-expressie; we richten ons op het embryo van het zebravis modelsysteem.

De ontwikkeling van een embryo kenmerkt zich door de vervolmaking van anatomische structuren in de tijd, van een eerste verschijning tijdens het ontwikkelingsproces tot en met het complete complexe orgaan. De expressie van genen in spatio-temporale patronen vormt de basis van het ontwikkelingsproces. Voor onderzoekers is een begrip van deze patronen in samenhang met de anatomische ontwikkeling belangrijk; hoe vormen de patronen de basis voor vorm verandering en welke genen kunnen bij dergelijke veranderende patronen betrokken zijn. In deze context hebben wij een omgeving ontwikkeld voor spatio-temporele gegevens uit embryonische studies van het zebravis modelsysteem. Een dergelijke omgeving is samengesteld uit verschillende componenten. Ieder component is afzonderlijk beschreven in de verschillende hoofdstukken van dit proefschrift.

Hoofdstuk 2 beschrijft een ontologie voor de anatomie van het zebravis modelsysteem. Een ontologie is een verzameling van concepten (termen) en relaties tussen die concepten. In een ontologie worden de relaties gebruikt om de concepten met verschillende manieren te benaderen, dit wordt granulatie genoemd. De ontologie voor de anatomie van het zebravis embryo kent verschillende groepen van concepten en relaties om de anatomie structureel te organiseren. Iedere anatomische structuur is een concept

type die beschreven wordt door andere concept typen (spatie, tijd en functie). Deze ontologie kan beschouwd worden als een taal die rekening houdt met functionele en spatio-temporele karakters van de anatomische structuren. Door de verschillende concepten en relaties kan een anatomische structuur worden benaderd vanuit verschillende niveaus. De ontologie voor de anatomie van het zebravis model organisme is georganiseerd in een database en wordt gebruikt op drie manieren; te weten (1) het exploreren van anatomische structuren en relaties (zie AnatomyOntology applet, hoofdstuk 2), (2) om beelden van de zebravis embryo te kunnen annoteren en (3) die beelden doorzoeken.

In hoofdstuk 3, wordt de 3D digitale atlas van het zebravis embryo beschreven. Deze atlas is een 3D referentiesysteem voor de anatomische ontwikkeling van het zebravis embryo. Tevens, wordt de atlas gebruikt als mal voor experimentele projectie. De atlas bevat een aantal 3D modellen in verschillende stadia van ontwikkeling. Ieder digitaal model is het resultaat van 3D reconstructies vanuit 2D histologische coupes; ieder coupe wordt gerepresenteerd door een digitaal microscopisch beeld waarin anatomische structuren zijn geannoteerd. Er wordt gebruik gemaakt van een semantische en een grafische annotatie om anatomische domeinen aan te tekenen. De grafische annotatie wordt gerealiseerd door een expert in het veld. De expert gebruikt een speciaal software (TDR-3DBase) om de contouren van anatomische domeinen te specificeren en de semantische termen toe te kenen. De anatomische termen uit de ontologie van het zebravis (hoofdstuk 2) worden gebruikt voor de semantische annotatie. Termen uit de ontologie geven aan atlas data een uniforme en een structurele annotatie. Deze structurele annotatie maakt het mogelijk om geannoteerde data toegankelijk te maken vanuit verschillende niveaus. De 3D atlas data, i.e. beelden en annotaties wordt opgeslagen in een database systeem. Een web-applicatie (AtlasBrowser applet) is ontwikkeld om de atlas database te doorzoeken. Concepten uit de ontologie worden gebruikt als zoektermen. Aan de hand van de zoektermen wordt een complete 3D embryo of onderdellen daaruit dynamisch samengesteld en gevisualiseerd.

Voor onderzoekers is een inzicht van genexpressie patronen belangrijk om de anatomische ontwikkeling te begrijpen. In deze context, een database is ontwikkeld voor

de opslag van patronen van genexpressie. 3D genexpressie patronen worden gerealiseerd door middel van (whole mount) *in situ* hybridisatie-experimenten die worden gevisualiseerd met behulp van een Confocale Laser Scanning Microscoop (CLSM). Op deze wijze worden 3D beelden verkregen. Deze database is een gereedschap voor onderzoekers om genexpressie patronen te kunnen analyseren en vergelijken. Daar toe is een "online" data submissie systeem ontwikkeld; hetzelfde kan worden gebruikt om deze data te doorzoeken en weer te geven. Bij de opslag wordt er gebruik gemaakt van de ontologie. Deze ontologie wordt gebruikt om spatio-temporele karakteristieken van genexpressie patronen te annoteren. Hierdoor wordt een uniforme en een structurele annotatie aan genexpressie data gerealiseerd. Deze vorm van annotatie maakt het mogelijk om genexpressie data te linken aan het 3D referentiesysteem (Atlas) en aan die van andere modelsystemen; als de ontologie bekend en beschikbaar is. Bij de data submissie zijn de metadata van de experimenten en data acquisitie geformaliseerd in een protocol. De ontologie in combinatie met dit protocol maakt de opslag en opvragen van data transparant voor de gebruikers. Het systeem voor patronen van genexpressie, i.e. GEMS wordt uitgebreid beschreven in hoofdstuk 4.

Data in de 3D atlas en het genexpressie database systeem zij geannoteerd met termen uit de zebravis ontologie. Deze ontologie maakt het dus mogelijk om data uit beide systemen te linken en op elkaar te projecteren. Daarnaast kan er direct worden *gelinkt* (verbonden) naar genomische data bestanden binnen het model en die van andere modellen. Om gegevens tussen de 3D atlas en genexpressie database te kunnen combineren is een systeem (query systeem) gerealiseerd dat gebruik maakt van het spatio-temporele karakter van de data. In dit systeem wordt 3D visualisatie gebruikt als visueel zoek interface om genexpressie te doorzoeken en te combineren. Een complete beschrijving van dit systeem is gegeven in hoofdstuk 5.

Naast het linken van onderzoek gegevens hebben we aandacht besteed aan het onderzoeken van verbanden binnen onze dataset; i.e. 3D patronen van genexpressie. Data-mining wordt toegepast om verbogen patronen in een dataset op te sporen voor om zo exploitatie en analyse maximaal te kunnen benutten. We hebben een aantal bekende data-mining algoritmen onderzocht en aangepast aan onze specifieke situatie in een case

studie waarmee we de principes van de analyse willen demonstreren. Een volledig beschrijving van de algoritmen en de resultaten zijn te vinden in hoofdstuk 6.

Publications

Bei, Y., Belmamoune, M. and Verbeek, F. J. “Ontology and image semantics in multimodal imaging: submission and retrieval”, Proc. of SPIE Internet Imaging VII, Vol. 6061, 60610C1 C12, 2006.

Belmamoune, M., Lindoorn, E. and Verbeek, F. J. 3D-VisQuS: A 3D Visual Query System integrating semantic and geometric models. In: InSCit2006 (Ed. Vicente P. Guerrero-Bote), Volume II "Current Research in Information Sciences and Technologies. Multidisciplinary approaches to global information systems", pp 401-405, 2006.

Belmamoune, M. and Verbeek, F. J. 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.

Belmamoune, M. and Verbeek, F.J. 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

Bei, Y., Dmitrieva, J., Belmamoune, M., Verbeek, F.J. Ontology Driven Image Search Engine. Proc. SPIE Vol. 6506, MultiMedia Content Access: Algorithms & Systems (Eds Hanjalic, A., Schettini, R., Sebe, N.), 65060G-1,65060G-10,2007

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

Richardson, M.K et al., with Bertens, L.F.M., Belmamoune, M., Verbeek, F.J. Zebrafish Developmental Patterning: New Tools for Medical Research International Journal of Developmental Biology, 2008 (In press)

Richardson, M.K. et al. with Belmamoune, M., Bertens, L.F.M., Verbeek, F.J. (2009) Zebrafish development and regeneration: new tools for biomedical research. Int. J. Dev. Biol. (2009) 53: 835-850.

Belmamoune, M., Bertens, L., Potikanond, D., Velde, R. v.d. and Verbeek., F. J. The 3D digital atlas of zebrafish: 3D models visualization through the Internet. (Submitted for publication, 2009).

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

Presentations at International Events

13th –16th July 2005 International European Zebrafish Meeting, Dresden, Germany The developmental anatomy ontology of the zebrafish: a common semantic framework for bioinformatics resources. Belmamoune, M., Bard, J., Welten, M., Verbeek, F.J.

6th October 2005 ICT Research Event, Eindhoven, The Netherlands. A digital zebrafish helps fishing for knowledge. Belmamoune, M., Bathoorn, R., Welten,M.C., Lindoorn, E., Richardson, M.K., Spaink, H. P., Siebes, A. and Verbeek, F.J.

15th-19th January 2006 San Jose Marriott and San Jose Convention Center, San Jose, California USA. Ontology and image semantics in multimodal imaging: submission and retrieval. Bei,Y., Belmamoune, M., Verbeek, F.J.

10th to 12th September 2007 International Workshop, University of Ghent, Belgium Developmental Anatomy Ontology of Zebrafish - An Integrative semantic framework Belmamoune, M., Verbeek, F.J.

20th to 22nd August 2008 International Symposium on Integrative Bioinformatics, Leucorea, Lutherstadt Wittenberg, Germany. Data Integration for Spatio-Temporal Patterns of Gene Expression of Zebrafish development: the GEMS database. Belmamoune, M., Verbeek, F.J.

Acknowledgements

This research project would not have been possible without the help and support of many people. I would like to take this opportunity to express my deep sense of gratitude to all of them. First of all, I would like to start with my thesis supervisor: Fons Verbeek who believed in me and gave me this wonderful opportunity to realize my childhood dream to do scientific research. Fons, it has been a great pleasure to do research with you and I hope this will continue.

Second, I would like to express my thanks to people within zebrafish bio-molecular informatics project, in particular Monique Welten, Ernst Lindoorn and Ronnie Bathoorn for their valuable collaboration. Especially Monique for the effort she put into providing me with her experimental results in the form of gene expression data and Ronnie for his helpful collaboration in the data mining part of my work. Special thanks goes also to all those that helped me by providing biological data, testing the applications or giving hints and solutions to the problems that I was confronted with. Here, I would like to mention explicitly Yun Bei, Laura Bertens, Michael Richardson, Joost Broekens, Hendrik Jan Hoogeboom and Jan Bot. All I have not specifically mentioned, I can assure that your help was valued very much and I sincerely thank you for your contributions.

Third, I gratefully acknowledge the support of several institutions. I thank N.W.O. BioMolecular Informatics research program (BMI) for the financial support and the Institute of Biology Leiden (IBL) for the unlimited collaboration to do research and to communicate with other biologists. Of course, I also owe my gratitude to the Leiden Institute of Advanced Computer Science (LIACS) and its staff for providing me with means including financial sponsoring to accomplish my research work in the most favorable conditions.

Fourth, I wish to express my gratitude to several persons whose support and help is not restricted to the scientific part of this project. I thank all my friends for being my friends. I thank all members of my beloved family, brothers and sisters and especially my parents, Aicha and El Fatmi Belmamoune; in my achievements I am grateful to their efforts and

unlimited support. I thank Bachir for believing in me from the beginning and Kebir for the stability that he gave when I just came to the Netherlands. I also wish to express my appreciation to Habiba, Bahija and Khadija for their assistance when they accepted to occasionally take care of my son Mahdi while I was busy with the last phase of my thesis. Their help was essential to complete this work. My special appreciation goes to the rest of my family, Rachid, Younes, Said and in particular Farida. Farida you are not only my sister but also my example and second mother. Thank you all for being constantly prepared to provide me with your help, encouragement and support.

Finally, I would like to thank my husband Rachid for his love, patience and interest. He shared the stress that I went through and he never complaint about the time that this work took from us. Rachid, living with you and Mahdi is a wonderful experience. Thank you and I love you….

Mounia,
October, 2009