Pattern Recognition in High-Throughput Zebrafish Imaging

Nezhinsky, A.E.

1 Introduction

Computers and computational methods have enabled experimental approaches in the life-sciences in which large amounts of specimens are processed in order to establish cause-effects relations or to detect rare events. These high volume experimental approaches are commonly referred to as High Throughput. In order to get the most out of such experiments, the aspects of experimental design and analysis need to be considered. Although computers enable High Throughput approaches, it is necessary to critically review the computational approaches that are addressed. Methods need to be adapted to the experimental design and sometimes, for several reasons, new methods need to be developed. High Throughput methods are applied in different fields of the life-sciences and the instrumentation for these High Throughput methods differs. In this thesis we consider High Throughput methods that are concerned with imaging, imaging with microscopes in particular. The scope of this thesis is to evaluate and develop methods for object recognition, extraction and measurement. In addition, given the output from a High Throughput experiment, the analysis of the measurement is addressed. The aim of this thesis is to find robust methods for object extraction and analysis. Throughout the thesis one particular application domain is used, infection studies in zebrafish for which High Throughput approaches are crucial. In order to understand the starting points for High Throughput methods, a survey of the literature is performed. For a review, two aspects have been particularly taken into consideration. The first being the relation between the experimental design and the aim of the experiment. The second being the computational aspects of the High Throughput experiment. In the next sections these two aspects are addressed. This introductory chapter closes with an overview of the contents of this thesis.

1.1 Whole Mount High Throughput Screening

A method called High Throughput screening (HTS) is used for scientific experimentation within different fields of modern biology. The method makes it possible to come to a high volume of experimental data. Results of such HTS experiments are used in setup for drug design and better understanding of various biological processes. There are different kinds of HTS methods, most often though they consist of typical basic steps which are applicable within different applications.

In order to prepare the experimental data a plastic micro-titer plate is used. Such a plate consists of a grid of wells (a multiple of 96). The wells are filled with the object of interest, which can be either an entire organism or a part of it. This is usually done by a robot in an automated fashion. The genetically altered object or organism is compared against a control group and is monitored through imaging. Resulting data is analyzed

in order to gain further insights in certain conditions. Experiments that are done for genomes do not involve the time factor, while experiments for proteomics testing are usually done in a time-lapse fashion. Depending on the speed of development of the organism, the start, end time and the interval of a time-lapse experiment is adapted, in such a way that the time resolution is sufficient. Nowadays HTS is made possible by automated microscopes. In order to do the final analysis problem specific image analysis software is being developed.

In [36] High Throughput (HT) methods are distinguished in 3 groups: Phenotype, Transcript and Proteomic. In this thesis we will focus only on the HT for phenotype analysis, since phenotype is directly related to imaging.

In order to be able to do in depth disease and infection analysis by tissue screening multiple models are used to mimic its behavior in the human body. A cell-based model is fairly inexpensive, it is, however considered physiologically distant to human. Vertebrate models of mammalians are more relevant, but expensive. A popular tradeoff between higher vertebrate models and cell based models are small organisms, like C. elegans, the fruitfly and the zebrafish. Because of their small size, these organisms can be imaged *whole mount*, which means the whole organism is treated and used in the imaging. We will discuss the HT possibilities of whole mount organism screening in more detail.

1.1.1 Non-Vertebrate Models

C. elegans (*Caenorhabditis elegans*) is an important non-vertebrate model organism that has a rapid development, is transparent and is easy to cultivate. The main advantage of the use of this model is the fact that C. elegans is the first organism of which the entire fate map from zygote to mature organism is available and it is the first organism with a completely sequenced genome. Fate maps provide a possibility to trace each cells predecessor through time during the development of the organism [56] (also known as cell lineage). The top of the fate map is the zygote cell followed at the next level by two daughter cells that at the next level divide into more daughter cells and in such a way a tree is formed and the nodes are indexed. This makes it very valuable in human disease exploration, since genetics of C. elegans are for about 70% similar to human. Computational aspects involve storing and traversing this cell lineage. AceDB (*AC. elegans DataBase*) is a database that is specifically developed for the C. elegans and can be used to browse the genomic data. Later AceDB was adapted to use with genomic databases of other organisms. Additionally C. elegans is widely used for its rapid development, transparency and ease of cultivation.

Rohde and Yanik [49] describe a HT approach for in vivo screening of C. elegans in a time-lapse fashion. C. elegans is immobilized by cooling it down. The proposed technology is used to perform subcellular-resolution femtosecond laser microsurgery of single neurons in vivo and to image the subsequent axonal (ALM/AWB) regeneration dynamics at sub-cellular resolution. Analysis of regeneration dynamics is done based on imaging and statistics.

1.1.2 Vertebrate Models

The Zebrafish (*danio rerio*) model is widely used as a vertebrate model. The layout of processes involved in human disease within the zebrafish organism often can be compared with their behavior in humans and therefore it can serve as a model for some human disease; an overview can be found in [2]. Zebrafish has a fast development and is easy to breed due to high fecundity rate: up to 200 offspring can be produced multiple times per week. It has a relatively small size which makes it easy to maintain large populations and produce large quantities for testing. The cell lineage of the zebrafish exists, but only in an early stage. The DNA of the zebrafish has a lot of homologs as compared to the humans, therefore the zebrafish model can be used to study the functional purpose of the genes [25] and is useful in the fields of developmental, genomics and drug research. It makes a good organism to use in research that is conducted in a HT fashion. Computational methods are needed to store genomic data and to browse it. In order to facilitate for HT data processing (of whole organism), computational approaches are needed for phenotype analysis, tracking and feature extraction.

Infection patterns can be tracked throughout the body due to transparency in early stages and especially when used with a fluorescent agent as a specific tissue promoter. Phenotypic development and infection pattern locations can be studied due to the external development of the zebrafish. In addition, evolving behavior of the zebrafish can be studied [20]. An early review of the HT approaches within the zebrafish research is given in [36].

Common carp (*Cyprinus carpio*) is also an important model organism due to it being a popular aqua cultivation species and its high reproduction rate. Most of the HT methods that are based on the carp model are about genomics and will not be considered in this thesis. However, nowadays carp is being considered in HT phenotype analysis.

Most HTS phenotype-based analysis methods are using brightfield or fluorescence microscopy.

Light microscopy based methods are used in order to do measurements on a whole mount organism phenotype for the aim of object localization and pattern recognition. A prerequisite would be analysis without the need to manually position the model organisms. One of the image based methods is Cognition Network Technology (CNT). CNT is a semantic and context-based segmentation approach that is applied in the detection of biological structures. In [62] CNT is used to detect transgenic fluorescent zebrafish embryos. The zebrafish embryos must be located in well plates and be presented in a lateral view. The shapes and regions can then be recognized by an artificial intelligence based algorithm that makes use of a predefined ruleset that use graded responses for area, length, and shape.

Another method for phenotype recognition for the zebrafish is presented by [34]. The recognition is based on a Support Vector Machine model that uses image descriptors (color histogram, representative color, and color layout) as input. This method does not perform detailed regional analysis but is defined as a three class classification problem. It is used to determine between three zebrafish embryonic conditions: hatched, unhatched, and dead.

Cellular processes within model organisms can be analyzed and modeled by making use of tissue-specific fluorescent reporter genes. Regions of interest can then be localized by image based methods that are able to retrieve fluorescent regions.

One of the methods for in vivo time lapse imaging is described in [3]. The authors describe protocols building a light-emitting diode fluorescence macroscope in order to screen adult fluorescent zebrafish. The macroscope is multi-spectral and imaging is being done simultaneously on multiple fluorescent channels. This method allows for the zebrafish to swim around. The method is able to distinguish up to five fluorescent proteins that are located within the fish and simultaneously image up to 30 adult animals that transgenically express a fluorescent protein. While the authors focus on the imaging part, the segmentation of the produced images is based on the different webcam inputs and thus limited to webcam capturing software (AMCap) and manual analysis. Aim of the study was to use this HT method for imaging transgenic animals, screening for tumor engraftment, and tagging individual fish for long-term analysis.

Walker et al. [64] describe a simple whole-organism HT method for quantifying changes in reporter intensity in individual zebrafish over time termed, Automated Reporter Quantification in vivo (ARQiv). ARQiv differs from current "high-content" (e.g., confocal imaging-based) whole-organism screening technologies by providing a purely quantitative data acquisition approach. Data processing was based on the comparison of intensity values within image regions. The values separated into the groups *background*, *signal* and *total signal* and comparison could be done across time in percentage terms.

Complex Object Parametric Analyzer and Sorter or COPAS is a combination of an instrument and a software package developed by *Union Biometrica* in order to analyze and sort biological objects that range from 100 to 1500 microns in diameter. Instrument is based on the principles of flow cytometry (cf. [35, 8]), but accommodates for larger objects and uses a pneumatic sorting mechanism. Originally COPAS was intended for the use with nematodes and later it was adapted for the use with zebrafish embryos. COPAS is used for HTS applications. In depth analysis of the produced images is done by the COPAS Profiler II software package. Based on fluorescence levels an optical profile is created. This facilitates in quantitative detection of multicolor fluorescent labels and determination of labeled object distributions such as tumor cells or granuloma in the zebrafish.

Saji et al. [19] describe a HT screening approach to assist in hazard ranking of engineered nanomaterials (ENM). This is done by tracking cellular injury in zebrafish embryos through epifluorescence microscopy. The analysis of the response parameters of nanomaterials is performed by the creation of heat maps by statistical approaches and the use of self-organizing maps (SOMs).

1.2 Pattern Recognition

The HT methods previously described all have significant ingredients of computer science. An overview of the described methods and their relationship to computer science (CS) is provided in Table 1.1.

ref.	exp. type	main CS component	v/t	output
[49]	f	regeneration dynamics by intensity, statistics	v	images
[62]	b	segmentation by ruleset, graded response	t	images
[34]	b	classification by SVM, image descriptors	v	image descriptors
[3]	f	segmentation by imaging of different channels	v	images
[64]	f	comparison by region based intensity, statistics	v	regional report
[8]	f	optical profile by intensity	v	images
[19]	f	response parameters analysis by SOM	t	heat maps

Table 1.1: An overview of some HTS methods. Experiment type (exp. type) is either fluorescent (f) or brightfield (b) and performed in vivo (v) or in vitro (t).

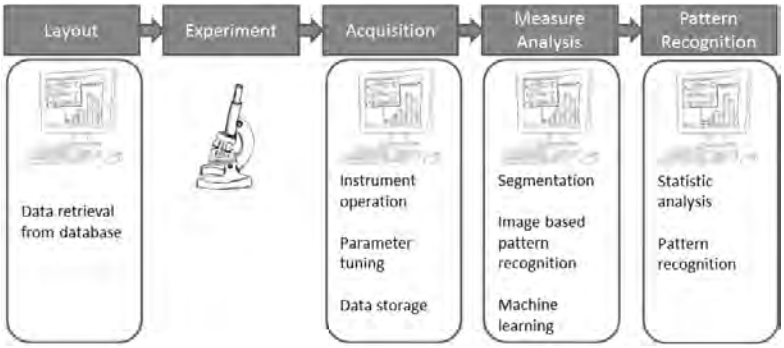

Figure 1.1: Schematic relations between a High Throughput Screening pipeline and Computer Science.

There are no standard methods which would be suitable for all applications and thus application dependent methods need to be developed. A schematic HT pipeline and its dependency on CS is shown in Figure 1.1.

Pattern recognition can be seen in two contexts: object recognition and data mining. Pattern recognition in terms of phenotypic object recognition plays an important role in almost every HT analysis process. Object recognition relates to looking for a pattern in the *image space*. In the case of separation of the objects of interest from the background we refer to it as segmentation. Data mining relates to looking for patterns in data *feature space*. Data mining depends on the data we use as input and can be used for finding global features followed by refining.

1.3 Outline of the Thesis

In this thesis we will in particular focus on two aspects of HTS: the image analysis step followed by pattern recognition of the resulting data. We have developed an image anal-

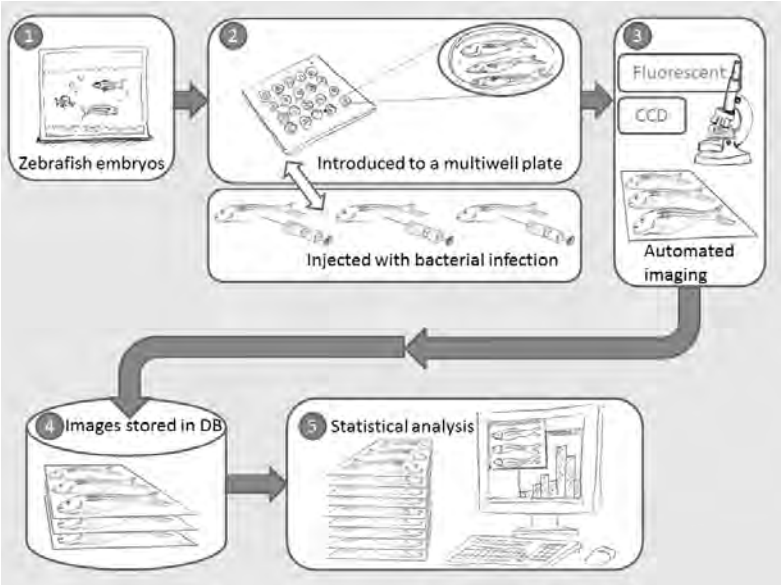

Figure 1.2: Example of a High Throughput Screening pipeline applied to zebrafish phenotype recognition.

ysis framework that is able to automatically recognize the zebrafish shape and retrieve the infection spread in a HT fashion. This pattern recognition framework allows for in depth analysis of the spread of infection within the zebrafish organism.

1.3.1 Image Processing and Segmentation

In the case study that is presented in this thesis we will be focusing on the zebrafish embryo as the model organism. We have developed a segmentation approach that is suitable for HT analysis and is able to retrieve the fish shape from images. An example of HTS is illustrated in Figure 1.2.

In Chapter 2 we describe our approach for retrieving the zebrafish shape from the input images. The approaches presented here are used not only to separate the foreground from the background but also to automatically recognize and annotate the zebrafish embryo shapes. We present three deformable template based approaches we have developed for this. In Chapter 4 we describe the graphical interface and the software that was created based on the algorithms described in Chapter 2 and Chapter 3.

1.3.2 Analysis of Statistics of Patterns

After retrieval of the zebrafish shapes and separating the input brightfield images into foreground and background the next step is to do measurements on the foreground area. In our case study the measurements are done on the response of the immune

system to the infection as well to the bacterial infection spread. In our research we aim to contribute to the understanding of the genes/proteins that are involved in the infection of the zebrafish with the Mycobacterium marinum, which is closely related to Mycobacterium tuberculosis, the causative agent of tuberculosis in humans. In chapter 5 we do a number of statistical assessments of the data that was collected by our HT approach. In Chapter 6 we discuss the results of the approach.