



Universiteit
Leiden
The Netherlands

Fuzzy systems and unsupervised computing: exploration of applications in biology

Cai, F.

Citation

Cai, F. (2017, December 12). *Fuzzy systems and unsupervised computing: exploration of applications in biology*. Retrieved from <https://hdl.handle.net/1887/57983>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/57983>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/57983> holds various files of this Leiden University dissertation.

Author: Cai, F.

Title: Fuzzy systems and unsupervised computing: exploration of applications in biology

Issue Date: 2017/12/12

Chapter 1

Introduction on Computation in Biology

As humans we have a good understanding of subjective concepts. For a computer, or simply a computational approach, it is much harder to deal with these subjective concepts as the boundaries in a subjective range are not well defined. In data analysis for life-sciences, we often encounter such problems and are challenged to find a solution that can deal with ranges of measurements in a robust manner.

In this thesis we particularly focus on finding solutions for data-analysis in the life-sciences. The life-sciences cover a broad field of research and approaches to deal with data-analysis which are typically multi-disciplinary. First one has to understand the particular field in of the life-sciences that the data-analysis is applied to and then, often, a number of techniques from statistics, mathematics, physics and computer science are employed to develop a solution. This multi-disciplinary approach for computational problems in the life-sciences is often captured under the umbrella of bio-informatics. One can state that bio-informatics is concerned with the analysis of data. As a consequence it is important to realize that the development of computational tools bio-informatics is therefore an implicit characteristic of this field. The consequence of working with experimental data and results from analysis is that these data need to be organized. These areas pretty much cover the field of bioinformatics.

Data from experimental set-ups in biological research are not always ideal for a straightforward analysis. Experimental conditions and biological variation both contribute to ambiguity. For analysis, the volume of data is not always sufficient, while the distribution of the data is uneven. Moreover, the measurement device itself, due to its electronic components, adds noise to the raw data. All of these issues have to be taken into account for an analysis. In order to further explore solutions for data analysis and typical for data sets without well-defined boundaries between its constituents, we investigate how the use of fuzzy systems theory can be used to enhance computations for such data sets.

Therefore, in this thesis we will explore the use of fuzzy systems theory for applications in bioinformatics. The theory of fuzzy systems is concerned with formulating decision problems in data sets that are ill-defined. It supports the transfer from a subjective human classification to a numerical scale. In this manner it affords the testing of hypothesis and separation of the classes in the data.

The fuzzy systems theory is part of the paradigm of soft computing, a collection of mathematical techniques that supports computing in dealing with uncertainty, inaccuracy, vagueness and incompleteness in data sets.

In the research presented in this thesis, we first formulate problems in terms of a fuzzy systems and then develop and test algorithms in terms of their performance with data from the domain of the life-sciences. From the results and the performance,

we will learn about the usefulness of fuzzy systems for the field, as well as the applicability to the kind of problems and practicality for the computation itself.

As mentioned, computing in bioinformatics is quite interdisciplinary; therefore we will use this introduction to present some of the major concepts from bioinformatics that are important to this thesis as well as provide scope of that field. Next, we will address soft-computing and in particular fuzzy systems and how this links with the analysis of data from the domain of biology. Further to this background information, we will explain our development of a heuristic-based pipe-line for data analysis, applied to generic data analysis as well as to bio-imaging.

1.1. The Scope of Bioinformatics: A Brief Introduction

The research field of Bioinformatics has matured over the past ten years and to date there is consensus on a definition. In general, bioinformatics is considered as the application of computational techniques of analyzing, managing and interpreting biological information [1]. The rationale is to create added value from the data for the field of biology [14]. The research field of bioinformatics encompasses a wide range of subjects, typically referred to as “omics” data, i.e. structural biology, genomics, proteomics, metabolomics, transcriptomics, cytomics, and image-based high-throughput studies.

In trying to understand biology, computational approaches have been probed. These were, in some cases essential for the understanding of phenomena, the discovery of inheritance by Mendel [2] in 1865 stands as a paradigm for computation in biology.

Modern approaches to computation in biology go hand in hand with the development and availability of computers. The notion of bioinformatics is developed in the late 1960's when molecular biologists started to compile their sequencing results of DNA and proteins in databases [3]. Initially, the field of bioinformatics was claimed by the research on the human genome but this progressed into the perception that bioinformatics had a much broader base.

The term bioinformatics is attributed to Hogeweg and Hesper [4], who coined the term as: “the study of information processing in biotic systems”. Over the past five decades, however, the field of bioinformatics has evolved in that it now involves various tasks, focusing on the analysis and understanding [14] biological data. Understanding refers to the creation of added value to the data.

Within the scope of bioinformatics different questions on biology are addressed. A common ground of all questions is that the starting point is a large amount of data from which understanding is developed in finding patterns in these data. This means that from the data, a systematic analysis is performed – these studies can be on

various levels; i.e. cellular behavior, molecular design and docking, metabolic networks, RNA/DNA and protein sequence alignment, RNA, DNA and protein structure prediction, analysis of gene expression data, just to name a few important fields.

Taking the data in a bioinformatics study as a starting point, the major emphases of bioinformatics become evident. Having large amounts of data requires these data to be organized in a structured form. In the early days of bioinformatics the important activity was constructing databases for the different data and making these databases available via the web. To date, this data management is still a major activity in the field of bioinformatics. Thus, developing databases and tools to archive and retrieve data is an important activity. The data are, of course, analyzed. A next major activity therefore is analyzing data and developing tools to achieve the analysis. From the analysis a higher level aggregation can be accomplished, combining results from analysis and finding patterns which contribute to the further understanding of biology. This activity, concerns working with statistics and machine learning approaches; it builds upon the other activities, however, its focus is to create added value from the large amounts of data in a manner that is meaningful to biology.

An important part of the field of bioinformatics is therefore the development of computational tools. Here there is common ground with computational biology where the emphasis is on theoretical models and simulations [14]. Nevertheless, computation and tooling is important to both fields that join forces in the quest of understanding the complexity in biology.

The crux for bioinformatics is to have adequate tools for analysis and interpretation available. There are ample computational approaches that have been successfully probed in bioinformatics studies and that, to date, are part of the algorithmic repertoire in bioinformatics. With different datatypes, different computational paradigms have been used. For analysis of sequences, i.e. RNA, DNA or protein, different alphabets are used in string matching procedures. The concept of dynamics programming has been very instrumental in being able to match strings in terms of their similarity. The Basic Local Alignment Search Tool (BLAST), to that respect is a major milestone for bioinformatics as a whole. Finding patterns from data is resolved using machine learning techniques; to a certain extent these techniques are inspired by biology, i.e. neural networks, genetic algorithms. In itself, machine learning techniques, for clustering and classification are deeply rooted in mathematics. Employing these techniques requires therefore, some understanding of the mathematics, e.g. choosing a fitting function in a classification problem.

Classification and clustering allows establishing relations in the data and to reason on behavior of biological entities. Techniques like Bayesian Clustering, Support

Vector Machines (SVM), neural networks and genetic algorithms represent forms of machine learning with and without control. It supports finding groups in the data as well as making predictions from new data based on prior classifications. Nowadays, a new revival of the concept of neural networks is embodied into so called convolutional neural nets, also known as deep learning. This is a very powerful machine learning approach that will further boost the understanding of biology from large amounts of data.

An important concept in computational approaches is that of heuristics, i.e. specific rules, so that a problem can be confined and computational pipelines for a specific bioinformatics study can be designed and implemented.

Our efforts for data analysis extend on existing approaches used in the field of Bioinformatics. Data from biological experiments do contain noise; and this noise complicates the analysis. Techniques that are based on heuristics are capable of confining the computation to solutions that are more probable. Heuristic-based techniques are therefore sometimes preferred in doing computations in large data volumes. Examples of such techniques are Bayesian nets, Neural networks, Fuzzy logic and evolutionary algorithms. In our effort to extend and improve data analysis in biology, we will focus on the so called fuzzy systems to see if we can reinforce solutions for datasets that are otherwise difficult to separate. Fuzzy systems are part of the soft-computing paradigm. Further explanation of this concept will be given in the next section.

1.2. Soft Computing and Fuzzy Systems

Soft computing [5], sometimes referred to as computational intelligence (CI), is a collection of methodologies that has become an area of formal study in computer science in the early 1990s. Soft computing differs from conventional computing in that it, specifically, exploits imprecision and tolerance in order to achieve tractability, robustness and low-cost solutions. The soft computing paradigm represents a number of techniques, its major constituents are the Fuzzy Systems (FS), the Rough Set (RS), Wavelets, Simulated Annealing (SA), the Support Vector Machine (SVM), the Artificial Neural Network (ANN), Evolutionary Algorithms (EAs) and Swarm Intelligence (SI).

The quality of the Fuzzy Systems is that it is rather easy to implement in a system ranging from very small and simple to embedding in large networked systems [17] [18]. As part of an analysis, the fuzzy systems will be part of a specific pipeline for data analysis. This is exactly how we intend to employ the application of the fuzzy systems in analysis of biological problems – and thus the data.

The concepts of **Fuzzy systems** have been applied to a range of different fields, from control theory to artificial intelligence, as well as to computational biology. The fuzzy systems are derived from fuzzy logic and it was first introduced by Lotfi Zadeh [6] in a monograph on fuzzy set theory in 1965.

In traditional computing, a system yields output(s) from the input(s), where conclusion is accepted to be either true or false. However, in real life situations, propositions are given with variable answers; for instance, degree of color between “yellow” and “red”, concept of “empty” and “full” in a water-filled bottle, sensation temperature of “cold”, “warm” and “hot” in a room, etc. In other words, we consider it natural to reason over a range of subjective concepts. If a certain concept cannot be defined exactly, an amount of quick and ambiguous definitions would develop. This typically happens in a group of people discussing a concept in a certain context. In Figure 1-1, this is exemplified, the fuzzy systems is the control methodology that mimics how a decision (description) is made by humans. Additionally, this decision-making process can be achieved and speeded up on the basis of prior experiences/recognitions of the individual.

Example:

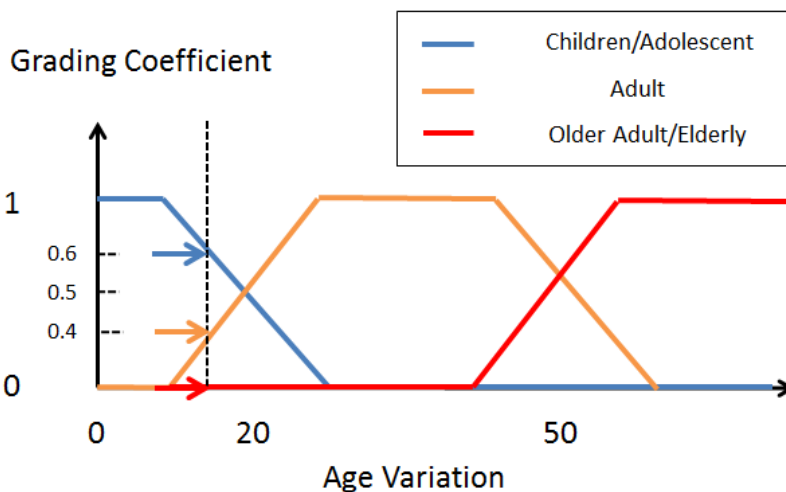


Figure 1-1. The application of Fuzzy Systems in the grading of human age. The grades of age are expressed as “Children/Adolescent”, “Adult” and “Older Adult/Elderly”; these are mapped by a pre-defined grading function onto an age scale. The age itself is ambiguous for human perspective inspection and/or sensible feeling. However, this sensation becomes obvious once a human-like metric is defined for the decision, e.g. when the grading coefficient is larger than 0.5, meaning a positive effect, vice versa. Since the red arrow points to zero, people within this age group (grading coefficient = 0, defined on Older Adult/Elder function) can be interpreted as “not old”; while the meanings of the age at orange (coefficient < 0.5, defined on Adult function) and blue (coefficient > 0.5, defined on Children/Adolescent function) arrow can be recognized as “fairly matured” but “still young”, simultaneously. Instead of concluding in either young (coefficient = 0) or old (coefficient = 1), fuzzy systems allows for decisions being made by users’ knowledge and experience accordingly.

The example of Figure 1-1 demonstrates two key ideas of the fuzzy system: first, the fuzzy system is able to model problems from concepts to mathematical paradigms that are based on the understanding and experience of the decision maker; second, the logic in fuzzy systems accepts the uncertainties that are inherited as realistic inputs, and thereby it is able to cope with these uncertainties (imprecisions) in such a way that their effects are negligible and henceforth, the system will result in a “precise”, human-like, output.

As mentioned before, fuzzy systems have been successfully applied to several areas, and also in bioinformatics. It helps in recognizing the hidden essentials in data by a degree of “truth” given by the fuzzy membership [19]. The fuzzy membership is a function that describes the weights for the contribution of the different levels in the system.

In this manner, using the fuzzy membership function, biological information is analyzed and interpreted based upon previous experiences [20] so that knowledge-based systems in biosciences are constructed by vagueness and uncertainty [21]. In this thesis, therefore, fuzzy systems based approaches are proposed and integrated into a dedicated data analysis pipeline(s).

Images form a particular class of data in life-sciences research. The data, i.e. the images, result from an imaging device and are sampled to a digital image. This digital image is input for a first data analysis in order to get measurements out of the image. The measurements themselves are input for a second analysis to find patterns over a collection of images. In general, this collection of images comprises an experiment. The data in the digital image are intensity values and these are ordered in a regular rectangular grid, directly related to the sensor in the digital camera. The data analysis is therefore completely adapted to this organization. Once we obtained the measurements, other approaches need to be probed. In this thesis, the image space based approaches, as well as for the feature space based approaches are

addressed; while the fuzzy systems approach has the capability to augment both the analysis. In order to get an idea of the scope, we briefly discuss data-analysis in bio-imaging in the next section.

1.3. Data Analysis in Bio-Imaging

The use of imaging techniques in biology is currently undergoing a revolution [7] with the availability of all new kinds of biological imaging techniques; i.e. fluorescence lifetime imaging, molecular imaging, diffusion-based imaging, X-ray based imaging, ultrasound-based imaging, magnetic resonance (MR) based imaging, etc. The in vitro and in vivo visualization of organisms, tissues, cells, proteins and macromolecular structures is enabled.

In the laboratory, bright-field and fluorescence imaging are routinely used; while the data from experiments is obtained by bio-imaging components. Analysis of data from these experiments can be performed rapidly by data-science oriented scientists, i.e. computer scientists and bioinformaticists. The data allows for data scientists to test and validate hypothesis related to a range of phenomena, e.g. cellular and molecular behavior, and the data can be acquired in different dimensions; i.e. 2D, 3D and time-lapse. Imaging techniques are part of the standard repertoire of a large range of experimentation with a visual control. However, the observations in the data, i.e. images, need to be verified and to that end computational tools are needed. Currently, large volumes of data are more and more the norm, therefore automation of the analysis is absolutely necessary. Such automation will result in data that support the interpretation of the experiment and are an indispensable extension to the imaging system (cf. Figure 1-2). Designing robust and accurate methods are being investigated thoroughly in biosciences and bioinformatics (cf. Section 1.1, [8]). With advances in technology leading to high-throughput systems for imaging, methodology design becomes increasingly important.

In this thesis, we consider data analysis as an extension of the imaging system aiming for understanding and recognizing information from biological image data. We distinguish a sequence of three major steps: (1) **image acquisition**, (2) **image processing**, and (3) **data analysis**. The acquisition is embedded in the imaging device which makes the image data available in digital form on a data repository. The image processing accomplishes a transformation of the raw image data to images from which reliable information can be extracted. In the data analysis step, observations are transformed to numbers and statistical representations so that analysis and validation can be applied. The final step in the analysis is to infer an interpretation from the measurements.

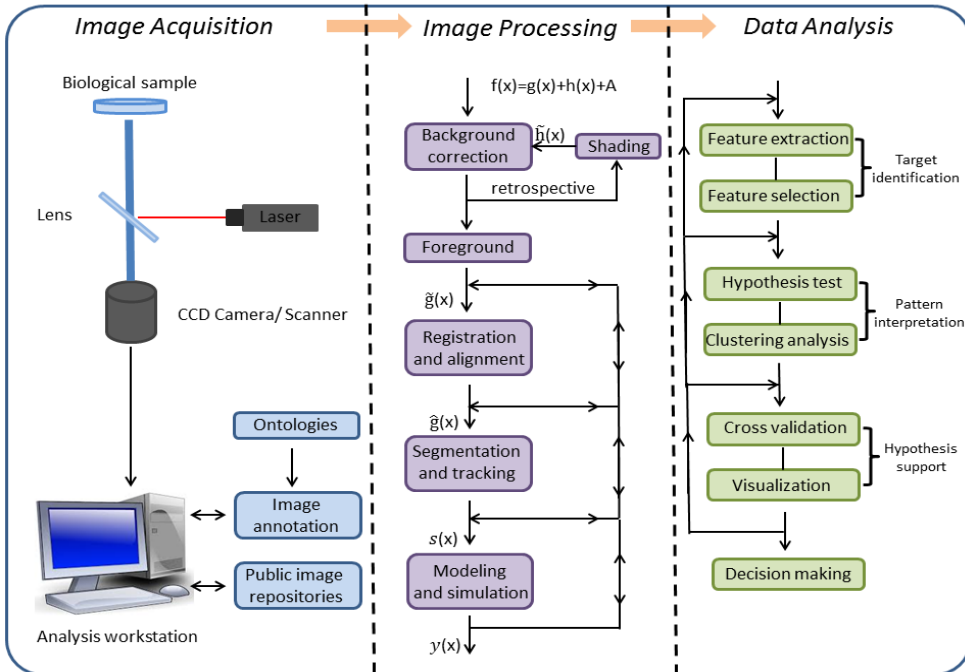


Figure 1-2. Workflow of data analysis in bright-field and fluorescence imaging system

Image acquisition. In bright-field and fluorescence imaging systems, image acquisition procedure is accomplished with a sensor array, often a CCD chip, which has a rectangular layout (CCD camera) or a line layout (flatbed scanner). A CCD camera is mounted on the optical system [9] whereas a line scanner is used to scan larger surfaces such as gels. In this thesis, we use bright-field microscopy images, i.e. cultured cartilage cells (cf. Chapter 2), fluorescence microscopy images, i.e. cardiomyocytes fluorescent images (cf. Chapter 2) and scanner images, gels of protein compositions from a range of cell lines (cf. Chapter 5).

For the acquisition of images, one should wish for the highest possible quality. However, there is a trade-off between image resolution and acquisition speed. A high-resolution detector allows imaging of objects whereas at lower resolution a significantly higher acquisition speed can be accomplished which is necessary to capture dynamics events. To further accelerate the dynamic acquisition numerous amount of efforts [10] [11] [12] have been made to attempt to acquire high-resolution images at high speed.

Acquired images will be stored in a repository. The increase of the data volume and complexity of biological experiments has made manual-workbook or generic databases unsuitable for keeping track of the images/data produced in experiments.

Therefore, image annotation, i.e. associating images with metadata, such as size, acquisition date, contents, is absolutely necessary for data provenance [22]. These metadata are often required for the further analysis to be able to understand an observation in context; or accomplish that an automated system can “understand” the context.

Image processing. In imaging images are the carrier of information and the content, at some point, should be transformed to a quantitative data presentation. Images as obtained from imaging systems are, however, far from “clean” (noisy) and need to be “polished”. The practical approach is to apply restoration through image processing before the data are being analyzed [15]. The term “image processing”, means to apply basic level operations on images. This is regarded to a pre-processing step such as enhancement, alignment and segmentation.

A well-performed image processing strategy, to some extent, can minimize data variation. It is, therefore, important to utilize empirical and problem-driven image processing solutions. Referring to Figure 1-2, a specimen is imaged and modeled by the input image $f(x)=g(x)+h(x)+A$, where true information $g(x)$ is masked by background noise $h(x)$ and all absolute multiplicative noise A . This raw image is then restored using an approach that employs additional images obtained at the time of image capturing, or through retrospective shading correction. For the resulting image $\tilde{g}(x)$, also known as the foreground image, now various other processing options are at hand: 1) a registration/alignment process producing $\hat{g}(x)$, 2) segmentation/tracking process, resulting in an image $s(x)$; or 3) an image modeling/simulating block with the output $y(x)$. After these image processing steps, information carried within experimental raw images is now enhanced and can be further explored.

Data analysis. From image processing we have obtained a restored image. Next, we extract features from the image. This is essentially a data reduction; we reduce the image elements to a set of measurements that sustain our observation. The analysis is in the heart of the bioinformatics methodology as it presents contextual approaches for data analysis, representation, and visualization. Image analysis deals with quantification of the amount and localization of signal, and measuring changes in structure over time. Data analysis can help to ensure that resulting measurements are accurate, objective and reproducible. Moreover, data analysis supports the further interpretation of the data by finding patterns in the data of an experiment or relating results to other experiments. In the context of biomedical research, this is typically the domain of bioinformatics. Commonly employed approaches involve target feature extraction and selection, data hypothesis test and data clustering, performance validation and visualization, as well as decision making. The decomposition and comparison of temporal biological data is not yet fully

understood [13]. In general, the data analysis in bioinformatics further augments the impact of knowledge discovery and making good predictions from the measurements.

1.4. Research Scope and Thesis Structure

In this thesis the capability of fuzzy systems is investigated with respect to systems in which simplifying an otherwise complex decision is augmented by allowing more hypotheses on the data. In the introduction (cf. Section 1.3), we have provided a dedicated bioinformatics data analysis pipeline, from which subjective and tedious image interpretations are alleviated. This thesis will further address the fuzzy systems in a number of different approaches on data measurement and develop an understanding on how the fuzzy systems can be employed in the concepts with pattern recognition.

The research described in this thesis is structured into 6 chapters. We have provided an introduction in **Chapter 1**. **Chapter 2, “Biological Image Background Correction”**, presents a strategy employing a combination of fuzzy logic and rough set theory to constrain a morphological image processing path during the process of image background correction.

Chapter 3, “Feature Selection Strategy in Region of Interest Mask”, illustrates a schema of feature selection via a fuzzy criterion in a multi-objective optimization algorithm. From this approach, sets of candidate solutions are provided to the researchers so that they can make decisions based on their own experiences/requirements;

Chapter 4, “Unsupervised Information Classification and Analysis”, elaborates on an unsupervised classification technique that hybridizes fuzzy uncertainty-based clustering method with swarm intelligence in order to find a good solution.

In these three chapters (cf. 2,3,4), the performance and efficiency of these algorithms are comprehensively assessed using various benchmark datasets that cover multiple facets of real-life situations. The results are compared with several commonly applied approaches; most are considered state-of-the-art methodologies. The evaluation and validation of these algorithms are used as a theoretical foundation for the design of image analysis workflows for experimental data.

Chapter 5, “A Systematic Study on One Dimensional Gel Electrophoresis Image Analysis”, employs the methodologies proposed in Chapter 2, Chapter 3 and Chapter 4 to interpret more complex biological problems in a multi-faceted manner. This case study presents a fuzzy-system based data analysis aiming to investigate and understand the identity of characteristics of proteins that are distinct/ shared

between different subgroups of cancer cell-lines. It further demonstrates the practical implication of image and data analysis workflows following the fuzzy-system designs, as promising.

Finally, in **Chapter 6** conclusions are presented and from a discussion an outlook to the further application of the fuzzy systems is further described.

1.5. Reference

- [1] Baldi, P., & Brunak, S. (2001). *Bioinformatics: the machine learning approach*. MIT press.
- [2] Ford, E. B. (1931). *Mendelism and evolution*. *Mendelism and evolution*.
- [3] Dayhoff, M.O. (1966) *Atlas of protein sequence and structure*. National Biomedical Research Foundation, 215 pp.
- [4] Hesper B, Hogeweg P (1970). "Bioinformatica: een werkconcept". 1 (6). Kameleon: 28–29.
- [5] Zadeh, L. A. (1994). Fuzzy logic, neural networks, and soft computing. *Communications of the ACM*, 37(3), 77-85.
- [6] Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3), 338-353.
- [7] Rittscher, J., Machiraju, R., & Wong, S. T. (2008). *Microscopic image analysis for life science applications*. Artech House.
- [8] Eils, R., & Athale, C. (2003). Computational imaging in cell biology. *The Journal of cell biology*, 161(3), 477-481.
- [9] Rochow, E. (2012). *An introduction to microscopy by means of light, electrons, x-rays, or ultrasound*. Springer Science & Business Media.
- [10] Picco, L. M., Bozec, L., Ulcinas, A., Engledew, D. J., Antognozzi, M., Horton, M. A., & Miles, M. J. (2006). Breaking the speed limit with atomic force microscopy. *Nanotechnology*, 18(4), 044030.
- [11] Oheim, M. (2007). High-throughput microscopy must re-invent the microscope rather than speed up its functions. *British journal of pharmacology*, 152(1), 1-4.
- [12] Bouchard, M. B., Voleti, V., Mendes, C. S., Lacefield, C., Grueber, W. B., Mann, R. S., ... & Hillman, E. M. (2015). Swept confocally-aligned planar excitation (SCAPE) microscopy for high-speed volumetric imaging of behaving organisms. *Nature photonics*, 9(2), 113-119.
- [13] Yan, K. (2013). *Image analysis and platform development for automated phenotyping in cytomics* (Doctoral dissertation, Department of Imaging and Bioinformatics, Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University).
- [14] F. J. Verbeek. (2016) Lecture Notes, "Bio Modelling and Petri Net".
- [15] F. J. Verbeek. (2017) Lecture Notes, "Image Analysis".
- [16] Jena, R. K., Aqel, M. M., Srivastava, P., & Mahanti, P. K. (2009). Soft computing methodologies in bioinformatics. *European Journal of Scientific Research*, 26(2), 189-203.
- [17] Liao, S. H. (2005). Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert systems with applications*, 28(1), 93-103.

- [18] Castillo, O., & Melin, P. (2014). A review on interval type-2 fuzzy logic applications in intelligent control. *Information Sciences*, 279, 615-631.
- [19] Cai F. & Verbeek F.J. (2015), Dam-based Rolling Ball with Fuzzy-Rough Constraints, a New Background Subtraction Algorithm for Image Analysis in Microscopy.. In: *Proceedings International Conference on Image Processing Theory, Tools and Applications (IPTA 2015)*. 298-303.
- [20] Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015). Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification. *PloS one*, 10(3), e0120364.
- [21] Jiang, Y., Chung, F. L., Ishibuchi, H., Deng, Z., & Wang, S. (2015). Multitask TSK fuzzy system modeling by mining intertask common hidden structure. *IEEE transactions on cybernetics*, 45(3), 534-547.
- [22] Vargas, E. L., Xia, Z., Slob, J., & Verbeek, F. J. (2016). A Semantic-Based Metadata Validation for an Automated High-Throughput Screening Workflow: Case Study in CytomicsDB. In *International Conference on Bioinformatics and Biomedical Engineering* (pp. 557-572). Springer, Cham.