Universidade do Minho
Escola de Engenharia

Ana Catarina de Jesus Domingues

Mining images of microbial communities for morphological characteristics in a support of clinical decision making

Outubro de 2013

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Ana Catarina de Jesus Domingues

Mining images of microbial communities for morphological characteristics in a support of clinical decision making

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho realizado sob orientação de

Professora Doutora Anália Lourenço
Professora Doutora Maria Olívia Pereira

Outubro de 2013

To my brother, António Domingues.

Acknowledgements

All my accomplishments in life are a result of my surroundings; therefore I am thankful for my dear friends, colleagues, and professors that made this journey possible.

To Professor Analia Lourenço for giving me this opportunity at the University of Vigo and for all the help and enlightenment during the past years. To Professor Maria Olivia Pereira, for her inexhaustible patience in answering my biological questions and giving me advice. To Ana Sousa, for the images and the biological insights. Professor Analia and Professor Olivia, you are an example to me.

I would also like to thank the Sing 33 group at University of Vigo, in particular Professor Florentino Riverola, Hector Vallejo, and Jeny Varela for making me feel at home. Thanks are also due to Nadine Santos, my partner in this adventure, for all the laughs.

I would like to dedicate this thesis to my brother António, my inspiration, my support. Thank you for challenging me to do this masters, and to not settle down in life. You are my mentor. Also, thanks to Andrea Freitag for all the love and support.

And finally, most importantly, my parents, António and Graciete. *Obrigada por todo o amor e apoio incondicional, sem vocês eu não era nada.*

Thank you all

Resumo

Um biofilme é uma comunidade de microrganismos envoltos por uma matriz extracelular produzida pelos próprios, que lhes garante proteção. Os biofilmes representam um problema para a saúde pública pois facilmente encontram-se em dispositivos médicos, podendo causar problemas graves para os pacientes.

Estudos prévios indicam algumas alterações observáveis do aspeto físico e bioquímico das comunidades microbianas na resposta à resistência e à virulência. Assim a morfologia da comunidade pode ser um indicativo da reação regulatória associada com fenómenos de patogenicidade microbiana.

O objetivo deste trabalho é por um lado a criação de um novo sistema de classificação de morfologia de colonia com medidas extraídas de softwares de imagem por outro lado, o estudo da classificação morfológica existente e do novo sistema de classificação, através de técnicas de mineração de dados com o objetivo de ajudar nestas classificações. Apresentamos vários softwares como solução que vão desde a caraterização da estrutura dos biofilmes até a caraterização de morfologia de colonia

Palavras-chave: Biofilmes, morfologia de colonia, processamento de imagens, mineração de dados

Abstract

Biofilms are communities of microorganisms embedded in a self-produced extracellular matrix, adherent to an inanimate, biotic surface that provides them with protection. Biofilms are a healthcare problem since they can be found in several medical devices and end up causing problems for the patients.

Previous studies have reported observable physical and biochemical changes of microbial communities associated with resistance and virulence response. This suggests that the morphology of a biofilm is a marker for regulatory interplays associated with the microbial phenomenon of pathogenicity.

The aims of this work are on the one hand, create a novel system of colony morphological classification with measurements extract from image software on the other hand study the current manual morphological classification and the novel one through data mining techniques. Here we present several software solutions to facilitate the process, from the determination of biofilm structure to the characterisation of colony morphology.

Keywords: biofilms, colony morphology, image processing, data mining

Contents

List of Figures

List of Tables

Acronyms

AR	Aspect Ratio
ASM	Angular Second Moment
cir	Circularity
CLSM	Confocal Laser Scanning Microscopes
CVF	Connected Volume Filtration
EPS	Extracellular Polymeric Substances
FISH	Fluorescence In Situ Hybridization
GFP	Green Fluorescent Protein
GLCM	Grey Level Co-Occurrence Matrix
IntDen	Integrated Density
PCC	Pair Cross-Correlation Function
Ra	Roundness Average
ROI	Region Of Interest
Round	Roundness
Rq	Root Mean Square
Rsk	Skewness
Rsu	Kurtosis
Stddev	Standard Deviation
XM, YM	Center of mass

Chapter 1. Introduction

1.1 Context and Motivation

Biofilms can be defined as aggregates of microorganisms embedded in a self-produced extracellular matrix and adhering to inanimate and biotic surfaces. Biofilms can be composed by one or more microbial species, but polymicrobial biofilms composed of several bacterial species are the most common [1]. Biofilms can also be formed by fungi microorganisms like *Candida albicans*. Besides the microorganisms, biofilms are also composed by interconnecting compounds which keep the microorganisms stuck to each other and to the surface. These compounds can be self-produced (such as polysaccharides, proteins, extracellular DNA and cell lysis products), substances derived from the immediate surrounding environment, or even dead cells [2].

Microorganisms within a biofilm can form long-term relationships, interacting with each other and establishing metabolic cooperation and/or antagonistic interactions. In a biofilm community, microorganisms tend to express different genes and proteins depending on the specific needs of particular biofilm region [3]. The genotype and phenotype alterations tend to be reflected in the morphology of the biofilm, i.e., the physical structure of the biofilm is somewhat reflective of how the biofilm cells interact with the environment [4]. Therefore, the capturing of images from different sections of the biofilm, with the corresponding quantification of the biofilm structure, is used to obtain insights about the biological processes that are taking place.

Furthermore, it is normal to culture microorganisms derived from biofilms onto solid media to characterize their growth patterns and to investigate their response to stimuli, such as their susceptibility profiles to antibiotic treatments. Culture onto solid media is a way to estimate the number of the biofilm-associated cells and their ability to grow. Theoretically, one viable biofilm-associated cell can give rise to a visible colony through multiplication. The morphology adopted by the biofilm-derived colonies formed on the solid media can also provide important insights about biofilm resistance, virulence and pathogenicity. For that reason, the studies related to the observable physical and biochemical traits of both microbial aggregates, i.e. biofilms and colonies, are often associated.

Colony morphotyping is an emerging topic of research due to its potential implication in antimicrobial resistance and increased microbial virulence. Indeed, the collection and characterisation of morphotypes of colonies such as the ones derived from multi-drug resistant pathogens, is envisioned as a key in supporting clinical decision-making. Therefore, any computational developments in support of automatic morphotype characterisation are seen as highly desirable and somewhat pressing.

1.2 Thesis contribution

Usually, the morphological features of a biofilm or a colony can be visualized with the help of a magnifier or a microscope. Depending on the specific characteristics of the aggregate under observation, different types of microscopes can be used, as well as different staining techniques. Despite this diversity, the comprehensiveness and accuracy of characterisation depends mostly on the expertise of the observer/annotator. Currently there are some software tools addressing the extraction of morphological features from biofilm-derived images, but this tools are usually designed to extract characteristics obtained under very specific conditions, namely produced under specific protocols.

A comprehensive evaluation of existing software is considered relevant as means to: (i) improve the computational tools at the researcher's disposal and (ii) standardize the characterisation of microcolony among the research community. It is also interesting to evaluate the viability of combining the measurements extracted by different computational tools to create a novel tool capable of classifying the morphological characteristics of a microcolony. Such a tool could be of tremendous value in (i) reducing the errors associated with the subjectivity inherent to human characterisation, (ii) reducing the time spent on this exhausting task, (iii) improving standards for morphological classification, and (iv) assisting research and clinical decision-making.

So far, morphotyping has been completely manual, relying on specialised curation. Yet, there is a portfolio of image processing tools that may be put into use here, with the benefit of alleviating manual curation as well as controlling annotation discrepancies (e.g., due to the degree of expertise of the curators). The goal is to seek a tool or combination of tools, and thus acquire a number of different morphological measurements. It would be important to equip researchers and clinicians with the means to reliably classify new images into well-studied colony morphology clusters. By doing so, therapeutic strategies and other decisions could be issued more promptly, assisted by the background of morphotype clusters meanwhile acquired.

1.3 Dissertation outline

The rest of this dissertation is structured as follows: chapter 2 provides an overview of related topics that were used as background to this work, while chapter 3 provides the study of the software available to since the biofilms structure analylis to the colony couting. In chapter 4 will be presented our study case. The results will be presented and discussed in chapter 5 and chapter 6 will closes this work with some conclusion and ideas to future work.

Chapter 2. Context

In this chapter, we aimed to go through the major concepts behind the subject being discussed in this dissertation. We address some key features like basic concepts behind biofilms, some of the observation techniques used in biofilms studies and a description of the morphological characteristics. It is also presented some image pre-processing techniques.

2.1 Biofilms

Microorganisms are the most successful form of life considering their habitats, their number, and their phylogenetic diversity. Microorganisms can exist in planktonic, in free suspension form, or associated in microcolonies. Biofilms can be considered a particular type of this latter form of living [1,3]. In fact, biofilms can be defined as communities of microorganisms immobilised on a solid surface and protected by a polymeric matrix produced by the microbial cells themselves [5,6].

These types of communities can be formed by one (single biofilms) or more microbial species (polymicrobial or mixed biofilms). In general, bacterial species predominate in biofilms, but fungi microorganisms, such as *C. albicans*, are often found in these living structures. Besides bacteria and fungi, algae and protozoa species can also be found in natural biofilms [7].

In a biofilm, the microorganisms are protected by a self-produced extracellular polymeric (EPS) matrix, which can have different densities and compositions[8]. This matrix may also encompass noncellular material such as mineral crystals, corrosion particles, clay, silt particles, blood components or cellular material, like extracellular DNA, proteins and cell lysis products. The presence of noncellular material depends on where the biofilm has developed. The diversity of the components that may exist in a biofilm makes the chemical analyses very challenging, especially in environmental biofilm samples [2,9].

The EPS matrix secreted by the biofilm-associated microorganisms protects them from hostile environments. Indeed, cells within a biofilm have a better chance of survival [10] than planktonic cells. Moreover, the matrix allows the cells to form long-term relationships with each other and to establish metabolic cooperation. The microorganisms that are closer to the surrounding environment appear to have an advantage, since they can easily acquire the metabolites needed, whereas those in the centre of the biofilm have more difficulty to obtain them [2].

The biofilm-associated microorganisms are different from the planktonic forms, because they can specialise, i.e., they can display special phenotypes and express different genes and proteins, depending on what is necessary in their biofilm region [1,3], as those involved in metabolism or starvation responses and in the reduced susceptibility of microorganisms.

As complex three-dimensional structures, biofilms may have internal channels through which nutrients and water can circulate. Due to high cell densities in the EPS matrix and limitations in the diffusion of metabolites, nutrient gradients arise readily in a biofilm community. Therefore, distinct chemical niches exist at different depths in biofilms. In fact, a biofilm often has areas with more oxygen than others, giving rise to aerobic and anaerobic zones at the same time [3,11].

Microbiologists have agreed on a model (see Figure 1) for the formation of biofilms. Biofilm establishment often starts with an attachment of planktonic cells to a surface, followed by the formation of cell clusters – microcolonies. Then, microcolony development and stabilisation occurs through the EPS matrix, which will provide protection from the surrounding environment. Bacterial cells can detach from the biofilm, due to, for example, lack of nutrients or high shear stresses. These bacterial cells can contaminate other surfaces, multiply and form a new biofilm in a more suitable environment [4,6].

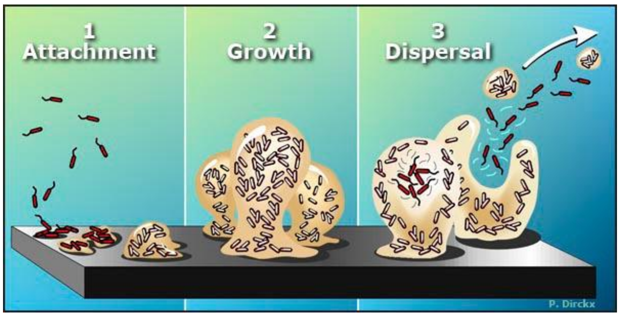


Figure 1: The Biofilm Life Cycle.
1. Planktonic bacteria encounter a submerged surface and within minutes can become attached. Cells begin to produce a slimy EPS matrix and to colonise the surface. 2. The EPS production allows the emerging biofilm community to develop a complex, three-dimensional structure that is influenced by a variety of environmental factors. Biofilm communities can develop within hours. 3. Biofilms can propagate through the detachment of small or large clumps of cells, or by a type of "seeding dispersal" that releases individual cells. Either type of detachment allows bacteria to attach to a surface or to a biofilm downstream of the original community [11]

Biofilms can form on just about any imaginable surface (Figure 2). In nature, biofilms can be encountered on hydrous solid or semi-solid surfaces, such as soil, rock material, animals, and plants. In areas related to human activities, biofilms may also be found on metals, plastics, kitchen counters, contact lenses, the walls of a hot tub or swimming pool, human tissue, indwelling medical devices, and industrial or potable water system piping. Indeed, wherever the combination of moisture, nutrients, and a surface exists, biofilms will likely be found as well [2,11].

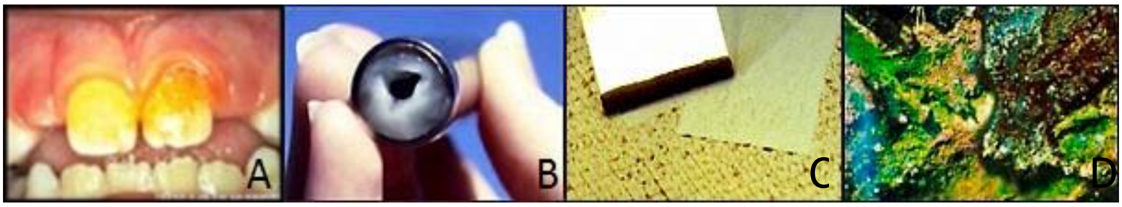


Figure 2: Biofilms on different surfaces.
A. Dental plaque is a biofilm B. Biofilm in pipe section C. Biofilm scraped from reverse osmosis membrane D. Biofilm in a stream in Yellowstone National Park [12].

Due to the several external and internal factors influencing biofilm formation, the biofilms can adopt a huge variety of sizes and shapes. Some of the most common biofilm structures are mushroom-shaped, pillar-shaped, or flat. Several authors claim that the morphological structure of a biofilm is influenced by the surrounding environment, i.e., the adhesion surface, the hydrodynamic conditions, the substrate available, and of course, the microorganisms that started the establishment of a biofilm as well as those further included in the biofilm consortia [2,11].

For example, in Heydorn, A., et al (2000), the authors describe the phenotypes of several biofilms formed by different bacterial species. *Pseudomonas putida* started with single cells on the substratum, and after growing into microcolonies, formed long filaments and elongated cell clusters. In turn, *Pseudomonas aeruginosa* colonised the entire substratum, and then formed flat, uniform biofilms. *Pseudomonas aureofaciens*, which is similar to *Pseudomonas aeruginosa,* had a stronger tendency to form micro-colonies. Finally, the biofilm structure of *Pseudomonas fluorescens* had a phenotype intermediate between those of *Pseudomonas putida* and *Pseudomonas aureofaciens*. They concluded that despite all the microorganisms described belonging to the family *Pseudomonadaceae* and being tested in the same experimental conditions, they form biofilms with different phenotypes[13].

Biofilms have an impact on human life, either directly by influencing human development, health, and disease, or indirectly by being involved in processes in natural or man-made environments [2]. This impact can be either beneficial or noxious. One of the beneficial effects of biofilms is in water treatment, where biofilm-associated microorganisms can degrade undesirable compounds and purify the water. On the other hand, biofilms can grow on ship hulls, causing increased friction and thus increasing energy consumption costs. Biofilm study is also important in industrial settings, as biofilms can develop inside industrial equipment, causing corrosion and equipment failure, which also results in increased costs [11].

There are several microbial aggregates in the human body, normally in mucous membranes and epithelial surfaces like the gastrointestinal tract, oral cavity, and skin. In normal conditions, the existence of these microorganisms is beneficial, for example degrading nutrients, synthesising vitamins, or helping the immune system [2].

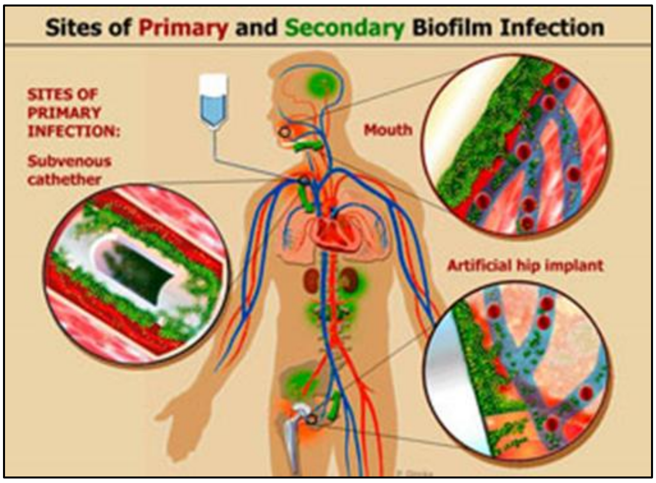

Sites of Primary and Secondary Biofilm Infection
SITES OF PRIMARY INFECTION:
Subvenous cathether
Mouth
Artificial hip implant

Figure 3: Common sites of biofilm infection in humans.
Once biofilms reach the bloodstream, they can spread to any moist surface of the human body [12].

The balance between the human body and the human body's microorganisms is complex. If the balance changes, this can result in infectious diseases. Biofilms also have a major importance in medicine (Figure 3), as they can be found, for instance, in medical devices, causing infections in patients, or in periodontal diseases, causing progressive destruction of the tooth-support tissues [14]. Moreover, biofilms are also related to infectious diseases, such as cystic fibrosis, where biofilms augment the severity of the disease [15]. These diseases can be caused either by members of the indigenous human microbial community or by microorganisms from the environment. If, for instance, the host is immunocompromised, injured, or suffering from cancer, harmful biofilms can develop in different organs and cause persistent infections. Bacteria which have been found to be involved in human biofilm-related infections are, for example, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Escherichia coli*, and *Dolosigranulum pigrum* [2].

As biofilms play a major role in human infections, are often encountered in biomaterial-related infections, and are associated with many nosocomial infections in medical units, one of the focuses of biofilms research is centred on biofilms that affect human health.

It is commonly accepted that the physical structure of biofilms determines how they interact with the environment (and the environment also determines the morphology of the biofilm). Mass-transport dynamics, hydrodynamics, and microbial community distribution are all factors known to influence biofilm structure. Also, as previously mentioned, the microorganisms in a biofilm are able to adapt to the surrounding environment by altering their gene expression that in turn affects the physical appearance of the biofilm [13]. Moreover, biofilms are always adapting, since they can be formed by several different microorganisms and the self-produced matrix can be different depending on several factors whereby the morphological structure is constantly altered. The study of the physiology and structure of bacterial biofilms is also important to understand their susceptibility to antibiotics. Therefore, it is important to study the morphology and spatial architecture of the biofilm in all these circumstances.

The diseases caused by polymicrobial biofilm infections are the most difficult to treat as they are often characterised by multiple and opportunistic pathogens whose interactions as a community increase the virulence [14]. The *in vivo* role of each microorganism in a mixed biofilm is still under discussion. The biofilm network is very complex and allows for mutual interactions that are just begun to be understood through *in vitro* studies.

It is normal in *in vitro* studies to culture microorganisms derived from biofilms onto solid media to characterise their growth patterns and to investigate their response to *stimuli*, such as their susceptibility profiles to antibiotic treatments. This type of culture is a way to estimate the number of the biofilm-associated cells and their ability to grow. Theoretically, one viable biofilm-associated cell can give rise to a visible colony through multiplication. The morphology adopted by the biofilm-derived colonies formed on the solid media can also provide important insights about biofilm resistance to antibiotics, virulence, and pathogenicity. For that reason, the studies related to the observable physical and biochemical traits of both microbial aggregates, i.e. biofilms and colonies, are often associated.

To study biofilms *in vitro,* it is necessary to identify and control the main factors that influence biofilm formation, such as flow rate, temperature, or nutrient composition. However, the development of bacterial biofilms is to a certain extent a stochastic process, and independent rounds of biofilm experiments do not always produce the same results even if the experimental conditions are kept constant. Therefore, it is necessary to make several replicas for the study to be valid. This fact increases the data produced in every study [13].

As mentioned above, a biofilm's genotype and phenotype is always changing. These constant changes tend to be reflected in the biofilm morphology. Under this assumption, researchers capture images from different sections of the biofilm and quantify its structure in order to obtain representative insights about the biological processes taking place.

The study of microbial communities is considered pivotal to help researchers understand how the social interplay between microorganisms enhances their ability to face and overcome environmental changes of various nature [16–23].

Physical and biochemical traits of microbial communities associated with resistance and virulence responses have also been reported [24]. Notably, the morphology of the community could be indicative of regulatory interplays associated with the microbial phenomenon of pathogenicity [24–27].

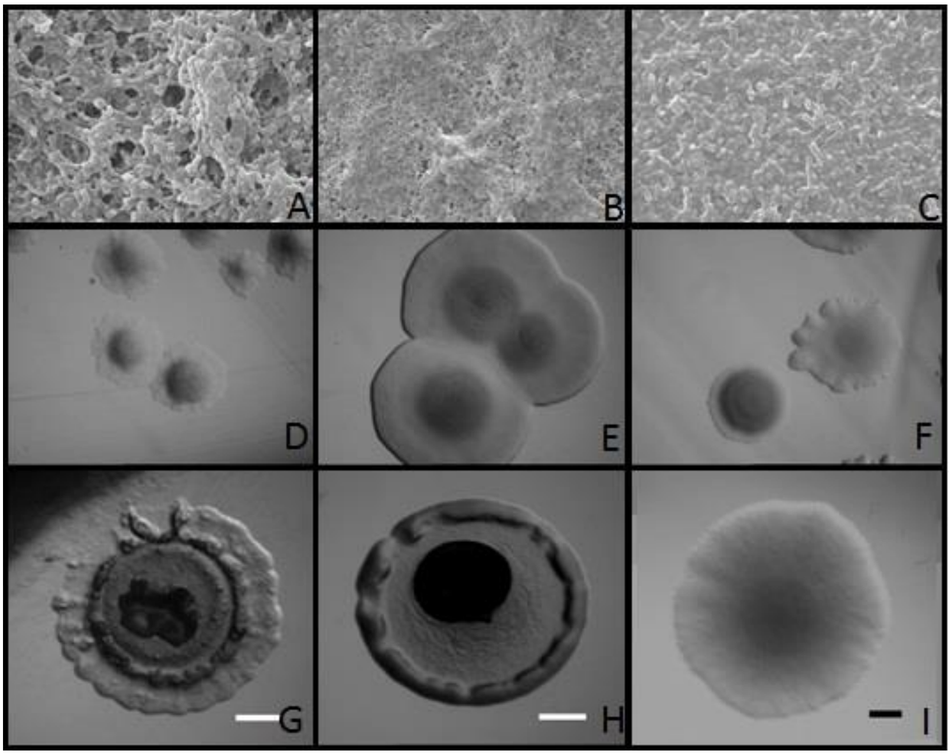


Figure 4: Several images relevant to the study of diseases associated with biofilms.
(A-C) Biofilm images obtained with a Scanning Electron Microscope (SEM). (D-F) Biofilm colonies. (G-I) Distinct biofilm colony for a more accurate inside view of a particular colony, obtained by a magnifier.

Figure 4 shows different types of images that can be studied to assess the degree of pathogenicity of biofilms. Images (A to C) have noticeably different characteristics. Taken with CLSM, this image type is normally used to study the structure of the biofilm. The images (D to F) are from an intermediate study. Researchers in the lab grow cultures in solid media and then quantify the number of colonies that grow in a specific period of time to assess the microorganism growth rate.

Afterwards colony morphology is often studied to evaluate the degree of pathogenicity. Images G to I are examples of a single colony.

In terms of decision-making, the ability to profile (at least, to a certain extent) the response expected from a community by observing its morphology could become a major breakthrough. Morphological measurements can be viewed as part of a signature. If such a signature is detailed in terms of genome and proteome, correlating particular morphological manifestations with regulatory responses, then one could query known signatures in assistance of decision-making. However, for these signatures to be a reliable reference, morphological characterisation should be comprehensive, accurate and unambiguous. The set of measurements to be considered should be well-established and described, and preferably quantifiable. Measurements should not depend on the ability of the researcher to describe the visual interpretation of the observation through common words [2,4,14,23,28,29].

Manual observation is a labour- and time-consuming task, and it is quite demanding in terms of expertise. The familiarity of the researcher with the type of microbial community under observation, namely the morphological switches of the organisms involved, is important. Technical skills, such as the ones related to image focus (i.e. the section of the community in display) and image quality (e.g. colour and resolution), are also required. Still, different interpretations of common morphological measurements may affect image annotation and further interpretation.

If, however, researchers could rely on data extraction tools for images to automatically characterise the morphology of microbial communities, expressing measurements such as size, form or roughness in numerical terms, the time consumed by the process would be reduced and the quality of image annotations would improve significantly. Image annotations could be effectively compared and thus, morphotyping could take a part in the decision making.

It is impracticable to analyse and classify every image produced in a single study by manual curation alone. Therefore, the use computational tools for this job is essential, thus also reducing the error potential and the time spent.

2.2 Microscopy - observation techniques

Biofilms are intriguing societies of microorganisms, and it is of general interest to unravel the processes involved in their development, physiology, and adaptation. However, due to their complexity, natural microbial communities have been challenging objects of investigation. In addition, biofilms are often located in places that are difficult to access, which makes direct and continuous examinations difficult. To reduce complexity and facilitate investigations in the laboratory under controlled and reproducible conditions, a number of biofilm model systems have been established. These include flow-cell-grown biofilms, colony biofilms, microtiter dish grown biofilms, and pellicle biofilms [30–32]. Combined with different staining techniques and different microscopes, these models help researchers acquire a better understanding of biofilms. In the following, this thesis will present some staining techniques as well as types of microscopy used in biofilms studies.

2.2.1 Staining Techniques

Biofilms are complex three-dimensional structures, which makes their analysis not trivial. While a single microorganism can be easily monitored using a conventional microscope, biofilms require, for example, additional resolution in the direction vertical to the substratum (the z-axis) [2].

Early biofilm studies by the Caldwell group employed a simple, yet efficient way of detecting the biomass in flow cells: the void volume, that is, the liquid phase, was supplemented with a solution of fluorescein iso-thiocyanate (FITC), leaving the biomass unstained. The resulting images were “negatives” and the biofilm could be rendered as the dark portions of the images. This gave sufficiently high resolution to determine, for example, cell sizes and spatial relations [5].

Staining techniques targeting the extracellular matrix such as lectins[1] or calcofluor white[2] can also be employed to visualise the surroundings of the biofilm cells. In addition, the extracellular DNA included in the matrix can be visualised by the use of different DNA-binding fluorophores[3]. Thus,

[1] Sugar-binding proteins

[2] Fluorescent stain that binds to structures containing cellulose

[3] Fluorescent chemical compound that can re-emit light upon light excitation

the employment of different staining techniques can help laboratories with fewer microscopy resources [2].

Syto series is one of the most used stain techniques. Syto series (Invitrogen, Carlsbad, CA) is a cell-impermeable dye with different excitations and emissions. These dyes are not harmful to the microorganisms, and can be used both in biofilms and microcolonies [33].

The combination of two types of stains has made the distinction of live and dead cells possible. The dye Syto 9 (green fluorescent) will stain all cells green regardless of whether they are dead or alive, while it is generally assumed that only cells with a damaged membrane will be stained by PI – propidium iodide dye (red fluorescent), indicating dead cells. Thus, the dead cells (cells with compromised membranes) will be stained red and the live cells (intact membranes) green [5,9,33].

Another way to distinguish live and dead cells is using the BacLight kit (Molecular Probes, Eugene, OR). In principle, bacteria that have been stained using the BacLight kit will result in red fragments for dead bacteria and green for live ones. Cells that contain both dyes appear yellow and should be treated as cells with damaged membranes. In case of computational tools, current quantification software treats co-localised pixels as both live and dead cells, thereby counting them twice during quantification. Visual distinction between green and yellow pixels can also be challenging [5].

The green fluorescent protein (GFP) has proven to be especially useful as a cell marker for ecological and environmental studies. GFP may also be used in order to investigate the protein location within bacterial cells. The applicability of various GFP types with different excitation and emission characteristics for specific labelling of different bacterial strains has been discussed in the literature. By combining GFP labelling of bacteria and Laser scanning microscope examination of the communities, major progress in the structure function of microbial biofilm systems has been achieved [9,13].

If genetic manipulation of the biofilm cells is possible, chromosomal tagging of cells with a gene cassette encoding the GFP can be a useful option. Alternatively, plasmids encoding for the GFP might be introduced into the cells prior to biofilm examinations. Depending on the construct, this fluorescent tagging can be used as simple labelling to verify the location of the cells in a biofilm, or, by selecting suitable variants of GFP genes and promoters, it can be used for monitoring gene expression in biofilms. Such tagging of biofilm cells has been done to monitor metabolic/physiological activity. Further, by using GFP variants with different emission spectra, such as the CFP (cyan fluorescent protein), YFP (yellow fluorescent protein), and RFP (red fluorescent protein), the spatial distribution of different species in a multi-species biofilm can be determined [2].

Another way of fluorescently labelling biofilm cells is through the use of fluorescent in situ hybridization – FISH, where specific probes hybridise to the 16S rRNA (Ribosomal RNA) in the cells. Because it involves probes with larger conjugates, this technique is preferentially applied on thin sections of thick biofilms. The number of ribosomes present in a given cell is proportional to the growth potential of the cell, and FISH labelling can consequently also be used to determine the growth status of a cell [2,14,34].

2.2.2 Microscopy

In the literature, the most used microscopes for the study of biofilms are confocal laser scanning microscopes (CLSM). CLSM is the method of choice for the monitoring of structure formation of living biofilms. As a result of its non-invasiveness and non-destructive character, CLSM enables the *in vivo* reconstitution of the three dimensional structure of microbial biofilms in their naturally hydrated form. CLSM can use a multi-channel modus where the different channels map individual biofilm components [35,36].

CLSM is an important method for the study of biofilm structure. Since its first application, CLSM has become widely used to improve the understanding of biofilm architecture. Multiple fluorescent channels can be recorded simultaneously, which offers the possibility to directly observe the development of individual biofilm components. Analysis of CLSM images has shown that biofilm communities form highly structured microbial assemblies. Studies using CLSM have further confirmed that the development of biofilms depends on various factors including mass transport properties, and have shown the importance of metabolic interactions within the microbial communities themselves [36].

The CLSM images can then be used for both qualitative and quantitative comparison and analysis. Confocal microscopy and derived methods require the specimen to be fluorescent. The biofilm must therefore either be auto fluorescent by means of indigenous fluorescent molecules, or the biofilm cells must express a fluorescent protein (e.g., GFP), or individual biofilm cells or other components of the multicellular structure must be stained [13,15].

The use of the confocal laser scanning microscope has helped overcome the apparent shortcomings of the conventional light microscope (the presence of out-of-focus light) by introducing point illumination and a pinhole, which allows optical sectioning of the specimen. The individual optical sections are subsequently assembled by aid of advanced computer software. Typically a biofilm with a thickness of more than 150 µm cannot be rendered with reasonable detail due to physical factors. The implementation of multiphoton excitation is a major step forward. Using a pulsed laser, it is possible to guide two (or more) photons to excite a fluorophore simultaneously. This means that the energy of the photons is combined to excite the target molecule. Using this technique, the depth resolution (i.e., the minimum distance to resolve two points) is increased manifold [2,14,15].

The successful analysis of microbiological samples with these advanced imaging techniques requires a number of considerations regarding the size and shape, preparation and mounting, necessity for probes, as well as the resolution and electromagnetic energy necessary for imaging and analysis. Ideally, the sample should be examined *in situ* in the fully hydrated state. This means that the fresh, living sample is directly used for imaging without chemical fixation. CLSM fully matches this necessity. In CLSM with an upright microscope, water-immersible (dipping) lenses proved to be ideal for imaging microbial communities. Restrictions in terms of sample size and mounting are the next issue. CLSM analyses only have restrictions in terms of the geometry (cm) of either the objective lens – microscope stage dimension. Another important point is the necessity for stains, fluorochromes and other probes. CLSM can take advantage of the intrinsic sample properties including reflection and autofluorescence. The photosynthetic pigments of algae and cyanobacteria are especially useful markers for differentiation of the two groups. If microorganisms can be labelled by reporter gene technology such as GFP or variations, then staining is not necessary. Nevertheless, in many cases, fluorochromes or fluor conjugated probes have to be applied for imaging of specific constituents and structures. This, of course, is a disadvantage as it may have an effect on the vitality of microorganisms. A further issue is the resolution at which the samples can be imaged and analysed. CLSM represents one of the most versatile tools for studying microbial biofilm systems. Its popularity is based on the current broad availability of CLSM instruments, the flexibility in terms of sample mounting and staining as well as the option for quantitative analysis of digital data. It is also a method of choice due to acquisition of three-dimensional data of the biofilm structure, used in a multi-channel modus where the different channels map individual biofilm components [2,36].

2.3. Image pre-processing

After obtaining an image of a biofilm or a microcolony that is going to be the object of computational analysis, it is necessary to pre-process it. Some of the visualisation software tools that are normally part of the microscope set-up have the tools to perform this pre-processing work.

The obtained images can be in different image file formats with different colour scales, sizes, etc. It is therefore necessary to process the image so that it is possible to apply the algorithms already developed. The most common, and normally the first methods to be applied, are those based on thresholding. Thresholding is a subjective operation, where the operator attempts to find the value on the grey scale that best represents the distinction between biomass and void space. There are biomass components that may be too transparent to be detected, which introduces some potential error into the measurements. Also, there is inherent error in the shadows and image noise that cannot be directly compensated for [37].

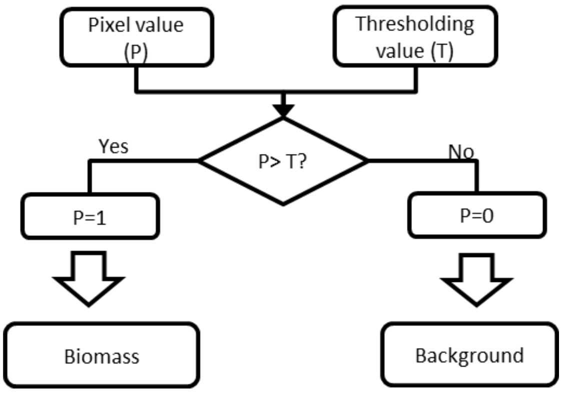


Figure 5 : Thresholding Algorithm Flowchart.
According to Comstat algorithm implementation, described in [13], if the pixel value is lower than the threshold value defined, the pixel is set to 1 which in this case denotes the biomass, otherwise the pixel value is set to 0, representing the background.

Although the algorithm is the same as the one described above (Figure 5) the thresholding algorithm defined by Yang et al., 2001 converts all pixel values lower than the threshold value into zero and all pixels values higher than the threshold value into one. Despite this discrepancy, the final image is similar. The only difference is whether the biomass is represented by the white pixels or the black ones.

The selection of a threshold level is therefore an important step in the quantitative analysis of an image of a biofilm. In fact, altering the threshold value will change the volume and morphology assigned to a given biofilm component. There is no consensus on the best method of thresholding nor on the best threshold value. In addition, no automated threshold procedure is guaranteed to work correctly with every image set since the characteristics of images from different samples, e.g. in terms of image histograms or spatial distribution of measurements within the samples, are widely changeable, so normally the user defines the threshold value [36].

The most described threshold method in the literature is the Otsu threshold. The Otsu threshold maximises the variance between the microorganism fluorescence and the background noise fluorescence, i.e. allowing for the separation of bacteria fluorescence from the background noise. This method does not constitute a significant computational burden for the image processing as a whole. This renders the method particularly suitable for image analysis systems, which will most likely be installed on personal computers [5,36]. Despite the Otsu threshold being the most widely used method, there are others mentioned in the literature, for example luminance thresholding where white pixels represent biomass and black pixels represent the background [14].

Following the binarisation of the image by thresholding methods, biofilm parameters are calculated from the binary image stacks, and a segmentation process known as connected volume filtration (CVF) is often performed [13]. CVF is a common method used to separate CLSM image pixels into connected biofilm and unconnected bacteria. After performing this algorithm, the bacteria that remain are the ones connected to the substratum (connected-biofilm bacteria). The bacteria eliminated in this algorithm are not connected to the substratum and presumed to be outside of the biofilm. The resulting matrix is a binary matrix where the connected-biofilm bacteria are represented. It is stored for calculations. This matrix is then used to quantify biofilm features, such as biomass, average thickness, roughness coefficient, and substratum coverage. Application of the CVF is optional, as users may prefer to include relevant floating material in their quantitative analysis depending on the characteristics of the system being analysed [5].

Segmentation comprises the two processes described previously: the thresholding process and the CVF process, and can be defined as the process of assigning pixels to distinct structural elements in the image, e.g., biomass, liquid media [35].

The other often used function is the pair cross-correlation function (PCC). This function quantifies the spatial arrangement patterns. The generated PCC curve allows for the determination of co-localisation, random distribution, or rejection (mutual avoidance) of two bacterial populations. This concept has successfully been applied to environmental biofilms and to in vitro-grown biofilm bacteria. The linear Dipole algorithm is also used to perform spatial arrangement analysis [14].

2.4 Morphological characteristics

Biofilm-associated organisms are able to adapt to environmental changes by altering their gene expression and general physiology, including increased resistance to antibiotics [38–44]. One of the ways in which microbial communities adjust to environmental changes is by changing the structural organisation of the biofilm [41,45,46]. Therefore, is necessary to proceed with a morphological characterisation of a biofilm. With the primary help of different staining techniques and CLSM, it is possible to achieve insight into the developmental process, spatial organisation, and function of a biofilm [2].

Numerous characteristics are used to describe the morphology of biofilms, which are then used in the development of software for biofilm image processing algorithms and tools. Each parameter measures a unique characteristic of either the cell cluster or interstitial space in the biofilm [37]. Depending on the number of dimensions considered, the parameters are divided into areal or textural for 2D images, or volumetric and textural in the case of 3D images. Textural parameters are calculated from greyscale images, and the areal/volumetric parameters are converted to binary images obtained after applying thresholding algorithms to the initial images. Areal parameters describe the morphological relationship between the size, and the shape of the surface measurements:

- Areal porosity, defined as the ratio of void area to total area.
- The average horizontal run length is the average number of consecutive pixels with a value of one (cell cluster) in a row (horizontal). Similarly, the average vertical run length is the

average number of consecutive pixels with value of 1 in a column (vertical). The average run lengths measure the expected dimension of a cluster of cells in each direction and are therefore a measure of the cluster size.

- The diffusion distance of a cluster is a measure of the distance (usually the Euclidean distance) from the cells in the cluster to the interstitial space. Diffusion distance is related to both the size of the clusters and their general shape. The diffusion distance is defined as the minimum distance from a cluster pixel to the nearest void pixel in an image, i.e. the minimum distance to a source of nutrients for the cell. A larger diffusion distance indicates a higher distance that the substrate has to diffuse in the cell cluster.
- Fractal geometry is used to quantify the roughness of an object. It is a mathematical system that allows objects to have a non-integral dimensionality, which is called the fractal dimension. In fractal geometry, the two-dimensional fractal dimension varies between 1 and 2. The higher the fractal dimension value, the more irregular the perimeter of the object. For the purposes of the analysis, the rougher the biofilm boundary, the higher the fractal dimension. For a more thorough description, see [37].
- Perimeter is the total number of pixels on the cluster boundary, which also relates to the accessibility of the nutrients [4].

From the grey level co-occurrence matrix (GLCM), it is possible to calculate the textural parameters [47]. Textural parameters have been less popular in quantifying biofilm structure to some extent because their relationship to biofilm processes is less intuitive. They measure the microscale heterogeneity in the biofilm by comparing the size, position, and/or orientation of the biofilm constituents:

- Textural entropy is a measure of randomness in the greyscale of the image. The higher the textural entropy, the more heterogeneous the image is.
- Energy measures the regularity in patterns of pixels and it is sensitive to the orientation of the pixel clusters and the similarity of their shapes. Smaller energy values mean frequent and repeated patterns of pixel clusters, and a higher energy means a more homogeneous image structure.
- Homogeneity measures the similarity of spatially close image structures: a higher homogeneity indicates a more homogeneous image structure. Homogeneity is normalised

with respect to the distance between changes in texture, but it is independent of the locations of the pixel clusters in the image [4,48].

- The angular second moment and inverse difference moment are similar measurements, but normalised for direction or distance respectively. Higher angular second moment values indicate more directional uniformity in the image, and inverse difference moment values indicate more or less variation in image contrast [47].

Volumetric parameters describe the morphology of the biomass in a biofilm. They are calculated with pixels representing biomass in the image. Each parameter quantifies a unique measure of the three dimensional image. In the literature, the parameters average run lengths, aspect ratio, diffusion distances and fractal dimension are also considered volumetric parameters. The other parameters are:

- Biovolume can be described as the number of biomass pixels in all images of a stack multiplied by the voxel size and divided by the substratum area of the image stack. The resulting value is biomass volume divided by substratum area. Biovolume represents the overall volume of the biofilm, and also provides an estimate of the biomass in the biofilm [4,13,36].
- The area of microbial colonisation defines the profiles of the fraction occupied by biofilm at the longitudinal plane, e.g. along the direction perpendicular to the solid substratum surface. This parameter can be related to the biofilm porosity profile.
- Colonisation fraction at the substratum, as the name suggests, is the fraction of the substratum surface colonised by the biofilm.
- Average height of microcolonies is the average height at which biofilm clusters rise from the solid substratum. This value is computed as the ratio between biovolume and the colonised substratum area.
- Interfacial area is measured as the area of the interface between voxels representing biofilm and those of the culture medium [35].
- Substratum coverage represents the fraction of pixels occupied by biofilm material for each image cross section. The fraction is defined as the ratio of foreground pixels to the total number of pixels for a given cross section and is then reported as a percentage.

- Area to volume ratio of an image stack is the number of foreground pixels which are connected to at least one neighbouring background pixel. The final value is then obtained by calculating the ratio area to volume ratio to biovolume[36].
- Thickness is calculated by a function that locates the highest point above each (x, y) pixel in the bottom layer containing biomass. Hence, thickness is defined as the maximum thickness over a given location, ignoring pores and voids inside the biofilm. The thickness distribution can be used to calculate a range of variables, including biofilm roughness. Mean biofilm thickness provides a measure of the spatial size of the biofilm and indicates the spatial dimensions of the biofilm.
- Roughness represents a measure of biofilm heterogeneity. The roughness coefficient is calculated from the thickness distribution of the biofilm. Biofilm roughness provides a measure of how much the thickness of the biofilm varies[13,36].
- Identification and area of distribution of microcolonies at the substratum is a function that locates microcolonies at the substratum, i.e. in the first image of the stack. Individual microcolonies are identified by 8-connected component labelling. Only microcolonies larger than a certain area size (determined by the user) are identified. The function calculates the total number of identified microcolonies, the area size of each microcolony and the mean microcolony area. The number and area sizes of microcolonies at the substratum provide valuable information about the organisation of the biofilm community. Substratum coverage reflects how efficiently the strain colonises the substratum.
- Surface to volume ratio is defined by the collection of pixels having at least one background pixel as a neighbour. In this case, the borders around the image stacks are all defined as biomass except for the top border, which is defined as background. In this way, only surfaces exposed to the nutrient flow are included in the surface area calculation. The surface to volume ratio reflects what fraction of the biofilm is in fact exposed to the nutrient flow, and thus may indicate how the biofilm adapts to the environment. For example, it could be speculated that in environments of low nutrient concentration, the surface to volume ratio would increase in order to optimise access to the limited supply of nutrients. Surface to volume ratio indicates how large a portion of the biofilm is exposed to the nutrient flow [35].
- The area occupied by bacteria in each layer is the fraction of the area occupied by biomass in each image of a stack. The substratum coverage is the area coverage in the first image

of the stack, i.e. at the substratum. Substratum coverage reflects how efficiently the substratum is colonised by bacteria of the population [13].

Chapter 3. Computational tools available for biofilm analysis

In many studies, such as [49,50] the analysis of CLSM data has been of qualitative rather than quantitative nature and consisted entirely of a visual image inspection. However, this approach is subjective, and not feasible when large quantities of data have to be analysed, which is often necessary to ensure the significance of the outcome of the analyses. For quantitative analysis of images of microorganism aggregates, computer software tools with different functionalities ranging from cell number counting to the classification of colonies morphotypes are currently available. Next, we will present a list (Table 1) several software tools for structural analysis that were reviewed.

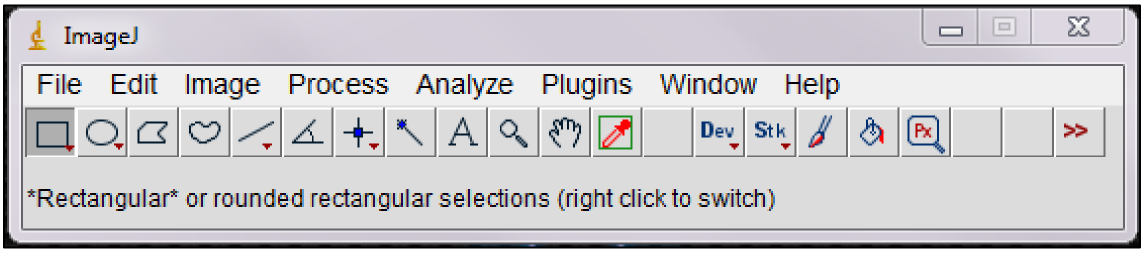
Table 1: Publicly available image analysis software tools for structural analysis of microorganism aggregates.
In the column "Characteristics" is presented the number of morphology biofilm characteristics described in the previously chapter

	Software licence	**Source code**	**Programing language**	**Operative system**	**Third-party dependencies**	**Graphical user interface**	**Biological context**	**Characteristics**
ImageJ [51,52] http://rsb.info.nih.gov/ij	Free	Available	Java	Linux OSX Windows	--	Yes	Generic	Not applicable
CellProfiler [53] http://www.cellprofiler.org	Free	Available	Python	Windows Linux Mac OS	--	Yes	Cell Yeast Colony	Not applicable
PHLIP- Phobia Laser scanning microscopy Imaging Processor [36] http://sourceforge.net/projects/phlip/	Free	Available	Matlab	Linux OSX Windows	Matlab license	Yes	Biofilms	10
DAIME- digital image analysis in microbial ecology [54] http://www.microbial-ecology.net/daime/daime	Free	Available	C++	Windows Linux	--	Yes	Biofilms	1
ISA3D - Image Structure Analyzer software [4,55] www.erc.montana.edu	Proprietary	Not Available	Matlab	Windows	Matlab license + toolbooxes		Biofilms	14
Comstat2 [13,56] http://www.comstat.dk/	Free(*)	Not Available	Java	Linux Windows	ImageJ	Yes	Biofilms	9
bioImage_L [33] http://bioimagel.com/	Free(*)	Not Available	Matlab	Windows	MCR MATLAB	Yes	Biofilms	2

3.1 Morphology description software

These programs are very different in terms of their runtime environment, file format acceptance, thresholding procedures, pixel/voxel/object recognition, subject of analysis, volumetric quantification, co-localization analysis, determination of structural parameters, and automation.

ImageJ (http://rsb.info.nih.gov/ij/) is a free, open source, and extensive scientific image processing software, developed by Wayne Rasban in 1987, and designed to handle various types of imaging data. Over the years it underwent several changes, keeping however the main ideas: (i) a biological image software that runs on any operative system, with the help of Java Runtime Environment; (ii) has a simple, user-friendly interface, with a single toolbar (see Figure 6); and (iii) extensibility via user-designed macros and plugins. Rasban chose a flexible approach to his software that allows the user to add functionality on their own, but in a manner that would allow sharing with others through macros and plug-ins. Macros are simple custom programming scripts that automate a task inside a large piece of software. The user does not need to have any programming skills to create a macro: in fact, with the help of a "macro record", it is possible to record any action manually, and thereby create a work flow that one can use repeatedly and share with others. Over 325 macros are currently available at http://rsbweb.nih.gov/ij/macros/. There are also over 500 plug-ins developed by users and available at http://rsbweb.nih.gov/ij/plugins/index.html. Since these plug-ins were developed to solve specific problems one can expect the continuous increase of this database. The plug-ins are designed to, for instance, count particles or enable the input of more specific instrument file formats – Bio-formats plug-in [51].



Rectangular or rounded rectangular selections (right click to switch)

Figure 6: ImageJ interface.

ImageJ is extremely versatile. It can display, edit, analyse, process, save, and print 8-bit, 16-bit, and 32-bit grayscale and 8-bit and 24-bit colour images. Image formats including TIFF, GIF, JPEG, and BMP can be imported and read as single images or stacks. ImageJ incorporates a number of useful tools for image processing. ImageJ can easily perform background subtraction routines and calculate the area, pixel value statistics, distances, and angles of user-defined selections (with several tools to select a Region of interest). It can also create density histograms and line profile plots. Standard image processing functions such as contrast enhancement, sharpening, smoothing, edge detection, and median filtering are supported as well [51,55,57]. The use of the ImageJ software is reported in [34] for the "Assessment of three-dimensional biofilm structure using an optical microscope". Hope and collaborators [58], used ImageJ to measure the thickness, and in, Barraud et al. [59] used it to calculate the percentage of the glass surface covered with biofilm.

CellProfiler (http://www.cellprofiler.org) is another good example of a generic yet, extensive image processing software. Due to the pipelining philosophy, it is possible to count colonies and classify them, for example according to the size, automatically identify objects, count them, and record a full spectrum of measurements for each object, including location within the image, size, shape, colour intensity, degree of correlation between colours, texture (smoothness), and number of neighbours [53].

The main purpose of bioImage_L (http://bioimagel.com/) is to allow easy interaction with the implemented image analysis tools, which primarily support input file preparation and output file display, as well as fast data pre-processing and processing, structural calculations of biofilm populations, and graphical displays of individual colour-based subpopulations with graphic outputs of the results. BioImage_L applies an in situ colour segmentation routine that automatically segments the colour image into individual pseudo-channels, and the areas and percentages of each identified colour subpopulation are calculated and presented. The principle of colour segmentation routine relies on the colour addition theory and classifies each pixel of the image into a predefined colour class, resulting in the generation of pseudo channels. Using one of the main advantages of CLSM biofilm images, the z-axis scans, it is possible to reconstruct 3D profiles, and using bioImage_L, calculate the surface and volume distribution of independent subpopulations of cells [4,33].

The Open Source software Daime (http://www.microbial-ecology.net/daime/daime) automatically recognizes 2D and 3D objects in single images and confocal image stacks, and offers special functions for quantifying microbial populations. Of note is the quantification of spatial localization patterns of microorganisms in complex samples like biofilms. It also offers many tools for analysing 2D and 3D microscopy datasets of microorganisms stained by FISH with rRNA-targeted probes or other fluorescence labelling techniques. The best quality of Daime is its visualization capabilities, which makes it possible to perform 3D visualization. Other features of this software are biofilm sectioning, spatial arrangement, abundance quantification, image segmentation and object editor [10,14,54].

Comstat (http://www.comstat.dk/) for flow cell biofilm microcolonies, PHLIP (http://sourceforge.net/projects/phlip/) for phototrophic biofilms, ISA-3D (www.erc.montana.edu) for structural as well as DAIME for fluorescence in situ hybridization (FISH) - stained sample analysis are examples of software tools that were especially designed to solve a specific problem in the lab group.

Xavier et al. (2003) developed software to quantify the area of microbial colonization profiles, biovolume, the colonization fraction at the substratum, the average height of microcolonies and interfacial area of morphology of a biofilm for single channel 3D image for CLSM [35]. PHLIP, Phobia Laser scanning microscopy Imaging Processor, extends the functions of the previous version of this software by including a new set of tools to perform quantitative analysis of large amounts of multichannel CLSM data in an automated way, a process necessary to produce statistically meaningful results [36]. PHLIP is also able to quantify ten different biofilm features: biovolume, substratum coverage, area to volume ratio, spatial spreading, mean thickness and roughness, fractal dimension in 2D and co-localization in 2D or 3D [2,33,36,60].

PHLIP was implemented as a MATLAB package and does not require additional toolboxes. The program was developed with flexibility and extensibility in mind, and its functionality can be easily expanded with new features. PHLIP therefore represents a platform for the integration of novel image processing operations without the need to code for import, export, or pre-processing functions. One example of the use of PHILIP to study biofilms is the description of the dynamic spatial and temporal separation of diatoms, bacteria and organic and inorganic matter during the shift from a bacteria-dominated to a diatom-dominated phototrophic biofilm [36].

The web-based PHLIP is a program that has a higher level of automation than the ISA and COMSTAT packages, also developed to quantitatively analyse single channel 3D CLSM data of biofilm imaging by determining a set of morphological parameters.

Similarly to PHLIP, ISA3D also has a previous version, called ISA - Image Structure Analyzer software [37]. It was initially developed for the UNIX/Motif environment in C++ with all calculations done in double precision arithmetic, and was able to analyse microscopy images of a biofilm. It was able to compute four areal and three textural parameters: porosity, run length, fractal dimension, diffusion distance and textural entropy, angular second moment inverse difference moment, respectively [2,55]. ISA3D on the other hand, was written using Matlab 7 and can calculate the same parameters as COMSTAT and Xavier et al.'s software. These parameters are: biovolume, volume to surface area ratio, porosity, surface area between biomass and voids, mean thickness, maximum thickness, and roughness coefficient. In addition, with the previous ISA capabilities, the user can now also analyse textural entropy, homogeneity, energy, areal porosity, average horizontal and vertical run lengths, diffusion distance, fractal dimension in two-dimensional image layers, and quantify their distributions with biofilm thickness. ISA3D is menu-controlled, user-friendly, and requires no prior knowledge in programming or image analysis but Matlab need to be equipped with the Image Processing Toolbox.

The measurements available in Comstat are: biovolume, area occupied by bacteria in each layer, thickness distribution and mean thickness, identification and area distribution of microcolonies at the substratum, volumes of microcolonies identified at the substratum, fractal dimension, roughness coefficient, distribution of diffusion distances, average and maximum diffusion distance and surface to volume ratio. All these measurements can be extracted from 3D stack of CLSM biofilm images [13,56].

Examples of analyses which could advantageously be made by Comstat include: (1) analysis of temporal structure development in single-species or community biofilms; (2) comparison of biofilm structures to different organisms or communities under steady-state conditions; (3) determination of the impact of specific mutations on biofilm structure; (4) analysis of the influence of different carbon sources or carbon source concentrations on temporal structure development in single-species or community biofilms; (5) analysis of the influence of antibiotic treatment on the biofilm structure [13].

3.2 Counting softwares

Research involving biofilm microorganisms often requires counting of bacteria colonies - an essential measurement in many widely used assays, such as biomedical assays. Bacterial colony counting is a low throughput, time-consuming, and labour-intensive process, since hundreds or thousands of colonies might exist in one Petri dish. Due to these difficulties, the process is often manually performed by well-trained technicians. There are also several automatic methods for the enumeration of bacterial colonies, Table 2.

Table 2: Publicly Available Image Analysis Software Tools for counting.

	Licence	**Source Code**	**Language**	**Operative System**	**Third-party dependencies**	**Graphical User Interface**	**Biological Type**
ImageJ [51,52] http://rsb.info.nih.gov/ij/	Free	Available	Java	Linux OSX Windows	None	Yes	Generic
CellProfiler [53] http://www.cellprofiler.org	Free	Available	Python	Windows Linux Mac OS	--	Yes	Cell Yeast Colony
OpenCFU [61] http://opencfu.sourceforge.net/	Free	Available	C++	Windows Linux	--	Yes	Colony
CellC – Cell Counting [62] https://sites.google.com/site/cellcsoftware/download	Free	Available	Matlab R2007a	Windows	--	Yes	Cell
Colony Counter [63] http://people.duke.edu/~kec30/colony.html	Free	Available	Matlab7	Windows	MCR MATLAB	Yes	Colony
CellNote [64] http://cellnote.up.pt/	Free(*)	Not Available	Java	Windows Linux Mac OS	--	Yes	Cell

(*)Open request

The number of software tools dedicated to colony enumeration has been increasing. One problem found in a quick research is that these software tools are linked to the hardware, i.e., the company that produces the microscopes/magnifiers also offers the software. However, with the global financial crises and the cuts in research funding, laboratories no longer have the funds to acquire them. Therefore, a free solution to reduce the time and the human labour is required. It is possible to find several (free) solutions in the literature, but they often need a special type of image and thus cannot be easily used, retrospectively, in images from studies not designed with these specifications in mind. To optimize the results for OpenCFU, for instance, (http://opencfu.sourceforge.net/) an image with the border of the petri dish is required [61].

CellNote is a novel software designed to count cells or subcellular structures in images and to help users to keep track of their annotated data, available at http://cellnote.up.pt/ [64]. Unfortunately, it was designed for cells, and its performance with colonies is not satisfactory.

There is a ImageJ plug-in, colony counter, for this problem http://rsb.info.nih.gov/ij/plugins/colony-counter.html [51,52,57]. This plug-in can improve the results with additional thresholding. Furthermore, it is possible to count colonies with the basic tool in ImageJ, “Analyse Particles”, after pre-processing a few with other tools, such as “Find edge”, to improve the results.

Colony Counter (http://people.duke.edu/~kec30/colony.html) is a hardware + software combination, but the software is available separately, and like the OpenCFU it needs the border of the petri dish [63].

Other software packages have their own limitations: CellC (https://sites.google.com/site/cellcsoftware/download) has difficulties handling outliers [62], and the well-known Cellprofiler (http://www.cellprofiler.org) makes it hard to configure a new pipeline.

Chapter 4. Computational morphological characterization of single microbial aggregates

As mentioned previously, it is normal when studying biofilms to determine their structural architecture as well as performing in vitro tests with microorganism derived from biofilms. One of the tests is to culture these microorganisms onto solid medium, with the purpose of characterizing their growth patterns and the appearance of the resulting colonies – colony morphology (see Figure 7).

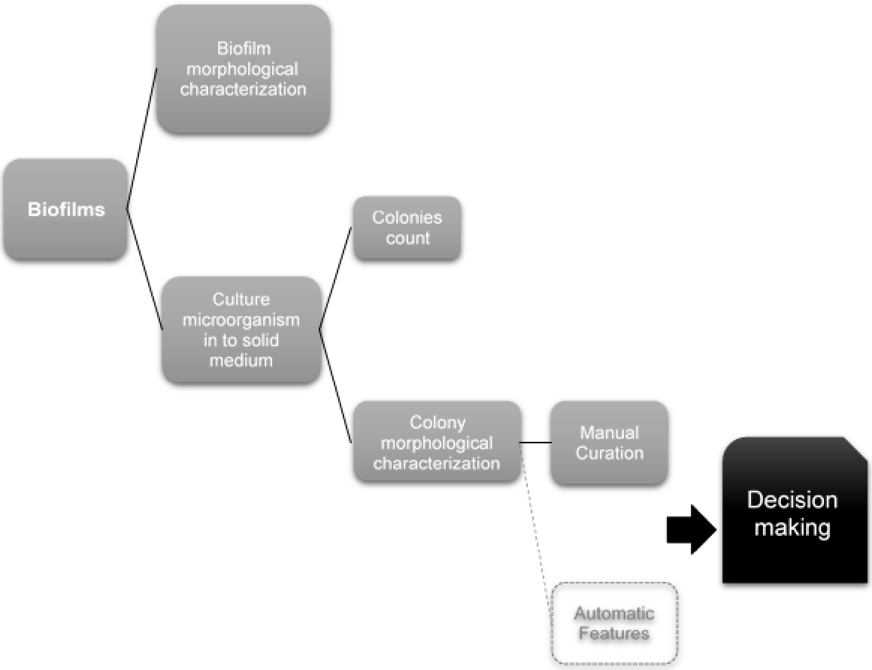

Biofilm morphological characterization
Biofilms
Colonies count
Culture microorganism in to solid medium
Colony morphological characterization
Manual Curation
Decision making
Automatic Features

Figure 7: Generic workflow of biofilm characterization with the end goal but not yet created of clinical decision making. A similar workflow was used to obtain the images that are used in this thesis.

Several studies report that changes in colony morphology are a sign of increased bacterial resistance to antimicrobial agents (i.e. antibiotics and disinfectants) and altered virulence and persistence [24,26,65,66]. It is therefore clinically relevant to study the colony morphology. Currently, the morphological characterization of a colony is performed manually by experts. This is a time consuming task and prone to inconsistencies due to the subjectivity inherent to qualitative measure performed by humans.

For these reasons, here it is presented the of manual curation and also the study of a possible group of measurements acquired with image processing software that try to morphological characterise morphologically the colonies. In this chapter we, also presented, an image processing workflow to acquire the measurements, using image processing software,

4.1 Description of the biological data

The images used in this thesis were previously curated manually by an expert of Biofilm Group. Its pathogenic morphotypes are previously described and publicly available on the online bacterial colony morphology database MorphoCol (http://morphocol.org/) [66]. The dataset consists of 111 images of the following colonies: 3 images of Escherichia coli colonies, 68 images of *Pseudomonas aeruginosa* colonies, 30 images of *Dolosigranulum pigrum* colonies and 10 images of *Staphylococcus aureus* colonies.

An Olympus SZ-CTV magnifier was used to enlarge and visualize the colonies, and the images were acquired via a CCD camera (AVC, D5CE; Sony, Tokio, Japan) and saved in the PNG format. The full image dataset is displayed in Figure 10.

The colony morphology used here is described in the Biofilms community Minimum Information about a Biofilm Experiment (MIABiE) portal (http://miabie.org/cmo.php). These guidelines consider two types of features to describe the morphology of a colony: the qualitative and quantitative (Figure 11). For this study quantitative features were ruled out, since was impossible to determine the age of the colony and the frequency purely by image analyses. Therefore only qualitative features were analysed, with the exception of colour of the colonies, as the images acquired are in greyscale.

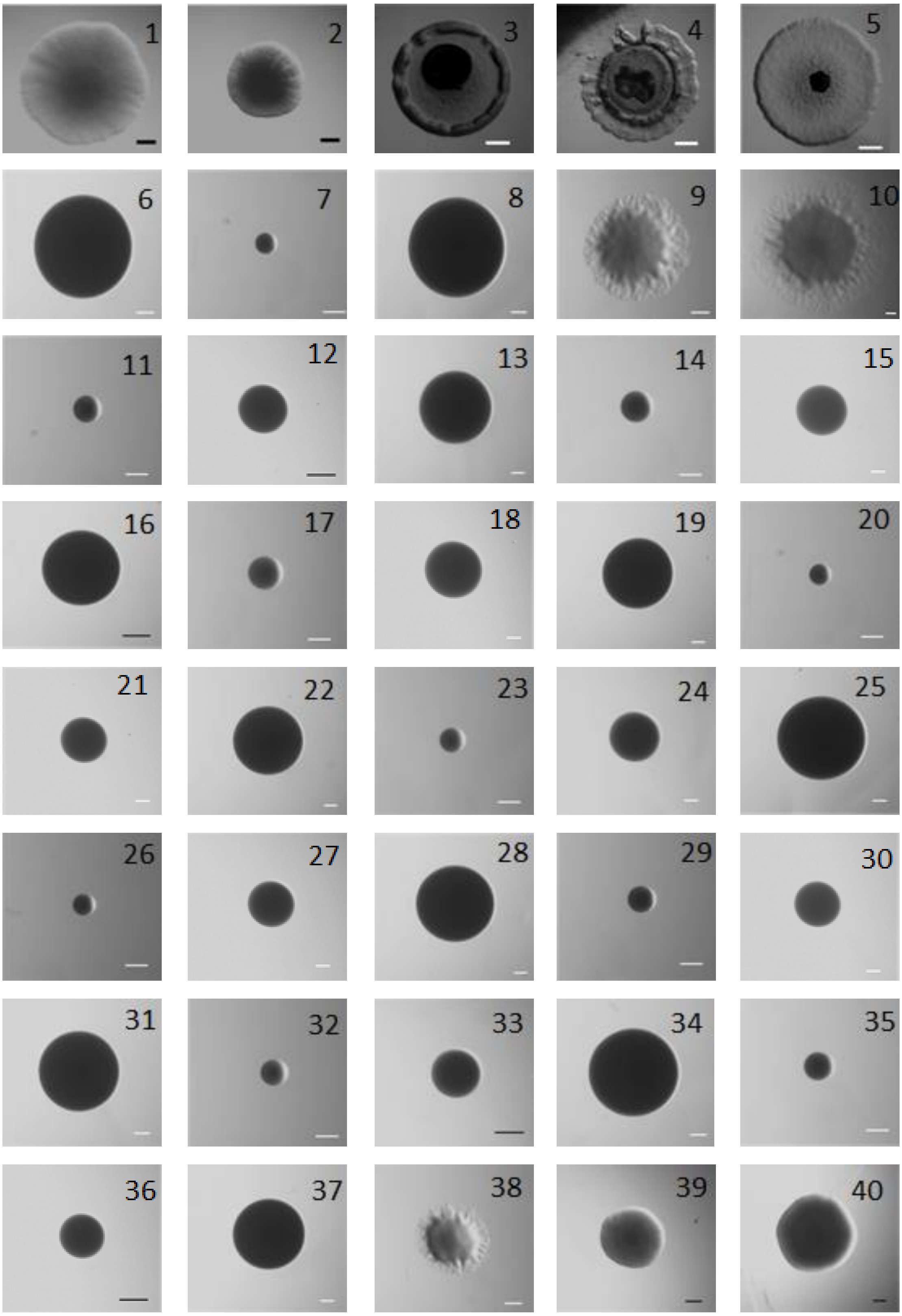


Figure 8: Full dataset of images used in this thesis work.

The scale of the image is represented by the bar in the image, the black bar represents 1mm and the white bar 0,5 mm (Cont.).

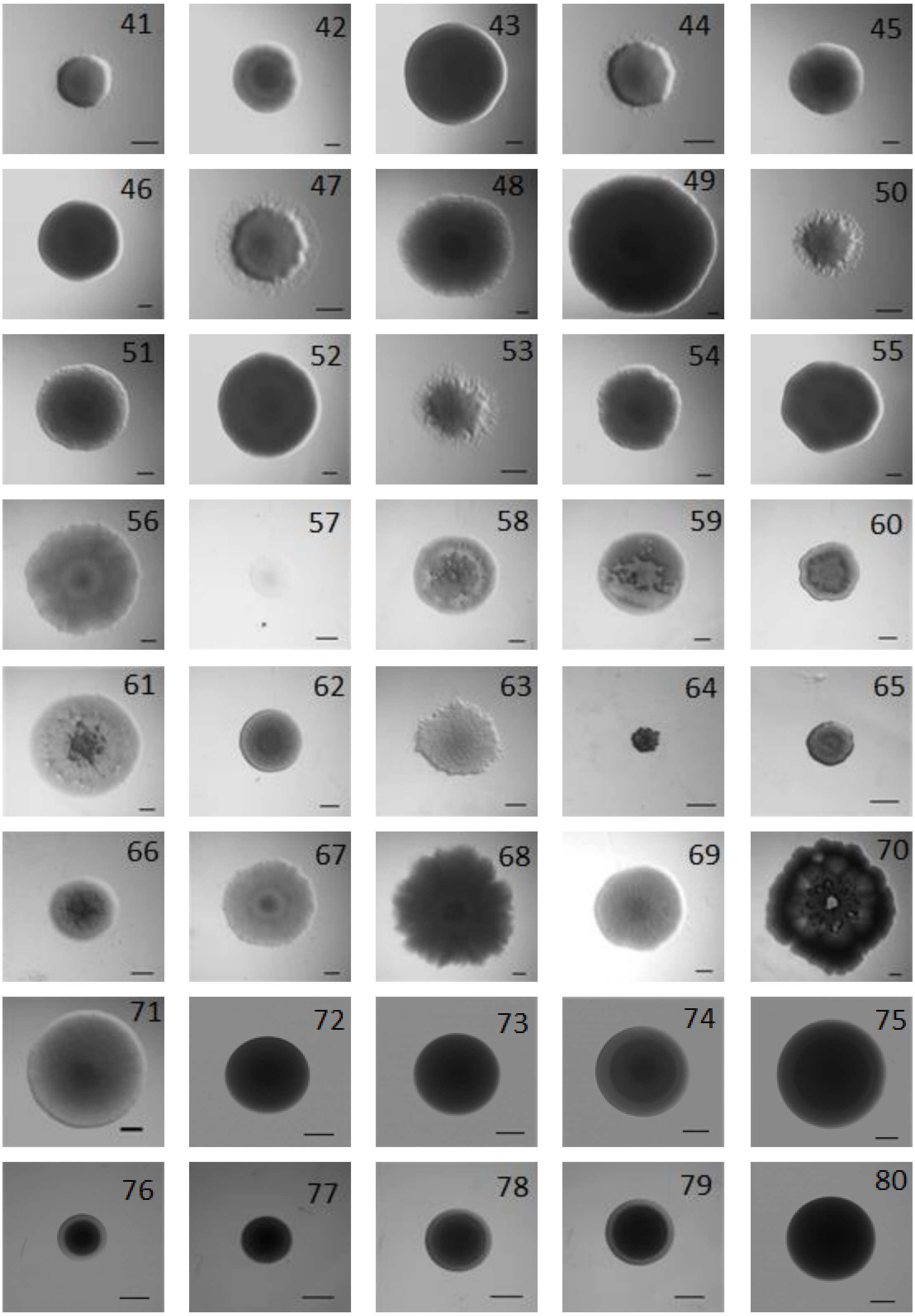


Figure 9: Full dataset of images used in this thesis work.

The scale of the image is represented by the bar in the image, the black bar represents 1mm and the white bar 0,5 mm (Cont.).

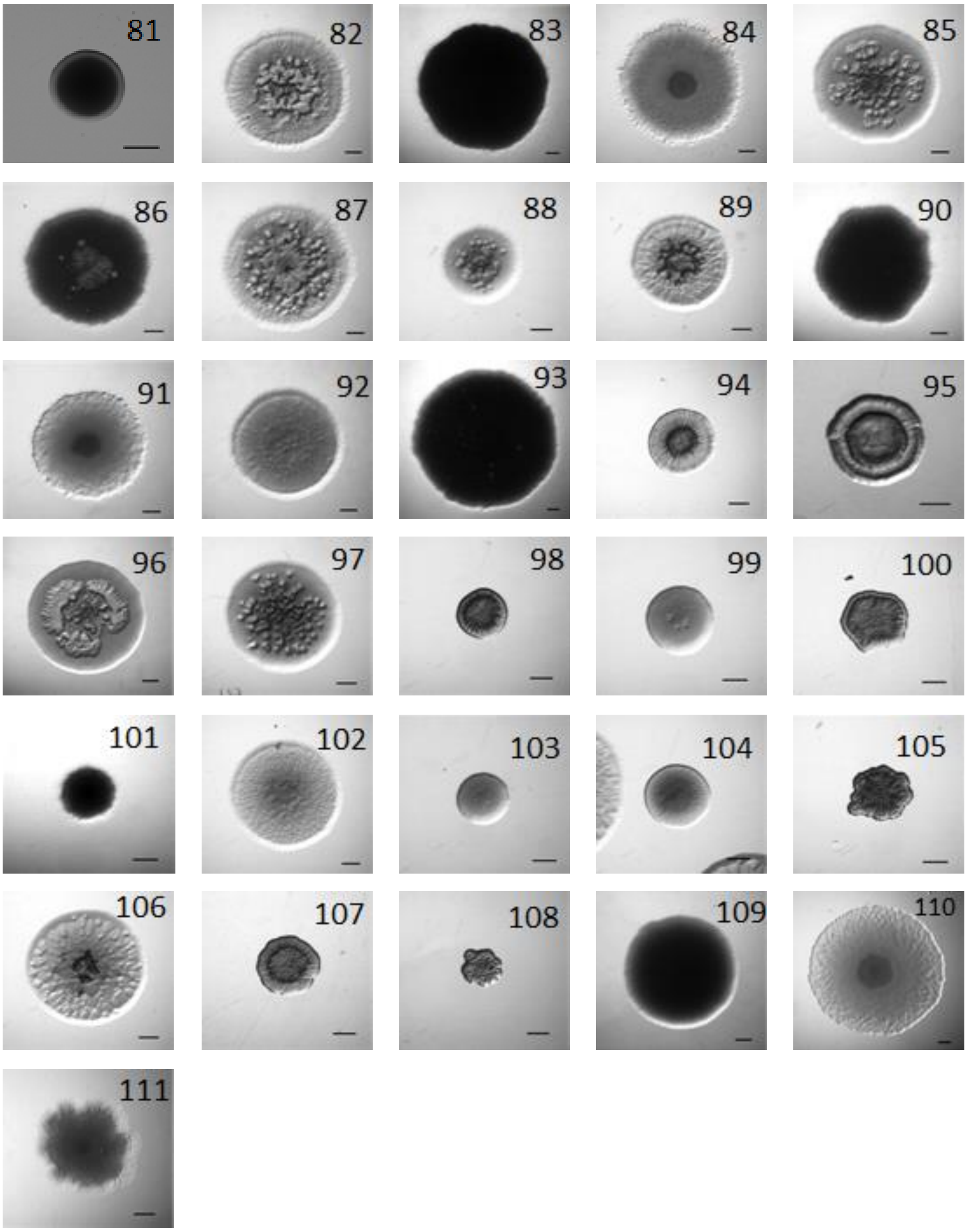


Figure 10: Full dataset of images used in this thesis work.
The scale of the image is represented by the bar in the image, the black bar represents 1mm and the white bar 0,5 mm(Cont.).

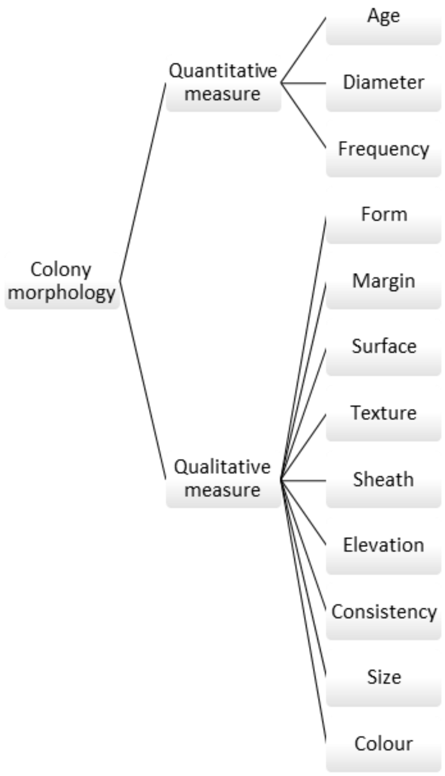


Figure 11: Colony Morphology according to http://miabie.org.
This morphology have to group of features, to this more the most important one is the qualitative measure group.

In summary, the manual image annotations considered cover the nine main morphological features: form, margin, surface, texture, sheath, opacity, elevation, consistency and size. The measurement units are nominal, based on common descriptions found in scientific literature. For example, the colony can be classified as erose, entire, irregular or undulate in terms of margin, and small or large regarding size. The manual classifications are present in Figure 12 and were based on the descriptions in http://morphocol.org/ and in MIABiE. Further information on the experimental setup available in the MorphoCol database and may be considered useful to understand some of the dissimilarities of annotation between computational tools and human observers. The manually curated annotations, for all images, are listed in Table 3, and will be referred to as “ManualC dataset” throughout the text.

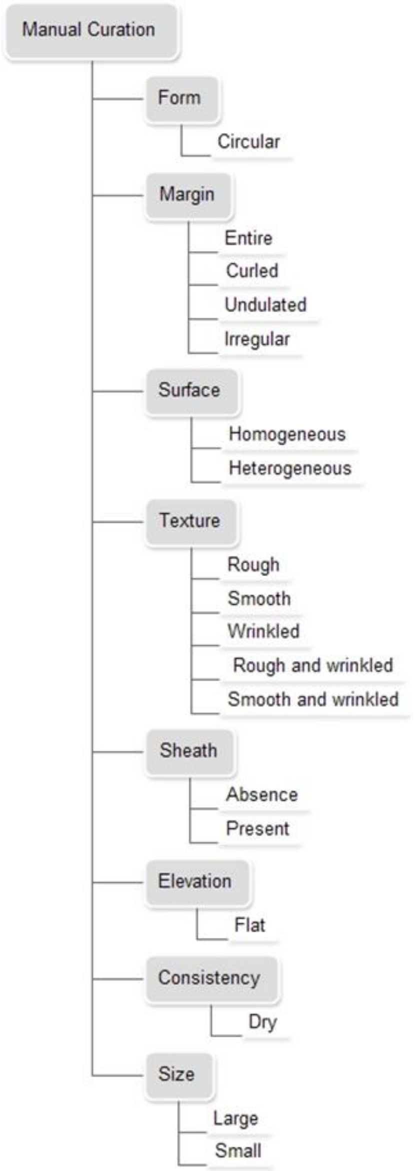


Figure 12: Detail of the manual curation measurements

Table 3: Classification of all the colonies used in this study according to manual curation measurements - ManualC dataset (see images in Figure 10).

ID	Form	Margin	Surface	Texture	Sheath	Opacity	Elevation	Consistency	Size
1	circular	entire	homogeneous	rough	absence	opaque	flat	dry	large
2	circular	entire	homogeneous	rough	absence	opaque	flat	dry	large
3	circular	entire	heterogeneous	rough and wrinkled	absence	opaque	flat	dry	small
4	circular	curled	homogeneous	wrinkled	absence	opaque	flat	dry	small
5	circular	undulate	heterogeneous	rough	absence	opaque	flat	dry	small
6	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
7	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
8	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
9	circular	irregular	homogeneous	rough	present	opaque	flat	dry	small
10	circular	irregular	homogeneous	rough	present	opaque	flat	dry	large
11	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
12	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small

ID	Form	Margin	Surface	Texture	Sheath	Opacity	Elevation	Consistency	Size
13	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
14	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
15	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
16	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
17	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
18	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
19	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
20	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
21	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
22	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
23	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
24	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
25	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
26	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
27	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
28	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
29	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
30	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
31	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
32	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
33	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
34	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
35	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
36	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
37	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
38	circular	lobulated	homogeneous	rough	present	opaque	flat	dry	small
39	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
40	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
41	circular	entire	homogeneous	smooth	present	opaque	flat	dry	small
42	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
43	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
44	circular	undulate	homogeneous	rough	present	opaque	flat	dry	small
45	circular	entire	homogeneous	rough	absence	opaque	flat	dry	large
46	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
47	circular	lobulated	homogeneous	rough	present	opaque	flat	dry	small
48	circular	undulate	homogeneous	rough	absence	opaque	flat	dry	large
49	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
50	circular	undulate	homogeneous	rough	present	opaque	flat	dry	small
51	circular	undulate	homogeneous	rough	present	opaque	flat	dry	large
52	circular	undulate	homogeneous	smooth	present	opaque	flat	dry	large
53	circular	lobulated	homogeneous	rough	present	opaque	flat	dry	small
54	circular	undulate	homogeneous	rough	absence	opaque	flat	dry	large
55	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
56	circular	undulate	homogeneous	rough	absence	opaque	flat	dry	large
57	circular	undulate	homogeneous	rough	absence	transparent	flat	dry	small
58	circular	entire	heterogeneous	smooth and wrinkled	absence	opaque	flat	dry	large
59	circular	entire	heterogeneous	smooth and wrinkled	absence	opaque	flat	dry	large
60	circular	undulate	heterogeneous	rough and wrinkled	absence	opaque	flat	dry	small
61	circular	entire	heterogeneous	rough and wrinkled	absence	opaque	flat	dry	large
62	circular	undulate	homogeneous	rough	absence	opaque	flat	dry	small
63	circular	erose	homogeneous	rough	absence	opaque	flat	dry	large
64	circular	erose	homogeneous	wrinkled	absence	opaque	flat	dry	small
65	circular	undulate	homogeneous	rough	absence	opaque	flat	dry	small
66	circular	undulate	heterogeneous	smooth and wrinkled	absence	opaque	flat	dry	small
67	circular	undulate	homogeneous	rough	present	opaque	flat	dry	large
68	circular	undulate	homogeneous	rough	present	opaque	flat	dry	large
69	circular	undulate	homogeneous	rough	absence	opaque	flat	dry	large
70	circular	undulate	heterogeneous	rough and wrinkled	present	opaque	flat	dry	large

ID	Form	Margin	Surface	Texture	Sheath	Opacity	Elevation	Consistency	Size
71	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
72	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
73	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
74	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
75	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
76	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
77	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
78	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
79	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
80	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
81	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	large
82	circular	undulate	heterogeneous	rough and wrinkled	absence	opaque	flat	dry	large
83	circular	undulate	homogeneous	smooth	present	opaque	flat	mucoid	large
84	circular	undulate	homogeneous	rough	present	opaque	flat	dry	large
85	circular	undulate	heterogeneous	smooth and wrinkled	absence	opaque	flat	dry	large
86	circular	undulate	homogeneous	rough	present	opaque	flat	dry	large
87	circular	undulate	heterogeneous	rough and wrinkled	absence	opaque	flat	dry	large
88	circular	undulate	heterogeneous	smooth and wrinkled	absence	opaque	flat	dry	small
89	circular	undulate	heterogeneous	rough and wrinkled	absence	opaque	flat	dry	large
90	circular	undulate	homogeneous	smooth	present	opaque	flat	mucoid	large
91	circular	undulate	homogeneous	rough	present	opaque	flat	dry	large
92	circular	undulate	homogeneous	rough	absence	opaque	flat	dry	large
93	circular	undulate	homogeneous	smooth	present	opaque	flat	mucoid	large
94	circular	entire	homogeneous	rough	absence	opaque	flat	dry	small
95	circular	undulate	heterogeneous	rough and wrinkled	absence	opaque	flat	dry	small
96	circular	undulate	heterogeneous	smooth and wrinkled	absence	opaque	flat	dry	large
97	circular	undulate	heterogeneous	smooth and wrinkled	absence	opaque	flat	dry	large
98	circular	entire	homogeneous	rough	absence	opaque	flat	dry	small
99	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
100	circular	undulate	homogeneous	rough	absence	opaque	flat	dry	small
101	circular	undulate	homogeneous	rough	present	opaque	flat	dry	small
102	circular	undulate	homogeneous	rough	absence	opaque	flat	dry	large
103	circular	entire	homogeneous	smooth	absence	opaque	flat	dry	small
104	circular	entire	homogeneous	rough	absence	opaque	flat	dry	small
105	irregular	undulate	homogeneous	rough	absence	opaque	flat	dry	small
106	circular	undulate	heterogeneous	rough and wrinkled	absence	opaque	flat	dry	large
107	circular	undulate	homogeneous	rough	absence	opaque	flat	dry	small
108	irregular	undulate	homogeneous	rough	absence	opaque	flat	dry	small
109	circular	undulate	homogeneous	rough	present	opaque	flat	dry	large
110	circular	undulate	homogeneous	rough	absence	opaque	flat	dry	large
111	irregular	undulate	homogeneous	rough	present	opaque	flat	dry	large

4.2 Computing morphological measurements

To decide which tool to choose for the comparison between manual and automated curation, we compared the measurements available in several image analyses tool (see Table 1), and in particular the modules whose output measurements are similar to the manual annotations. Both commercial and freely available software was compared, and there is a set of measurements common to most of them. Notably, ImageJ the free Java coded image tool developed at the National

Institutes of Health provides for them all [52]. Thus, this was the tool of choice for all image analyses.

4.2.1. Image pre-processing

Before proceeding with the extraction of measurements from the image, it was necessary to perform a few pre-processing steps presented in Figure 13 to improve the quality of the measurements extract and also to find the region of interest, i.e., the colony, in the image.

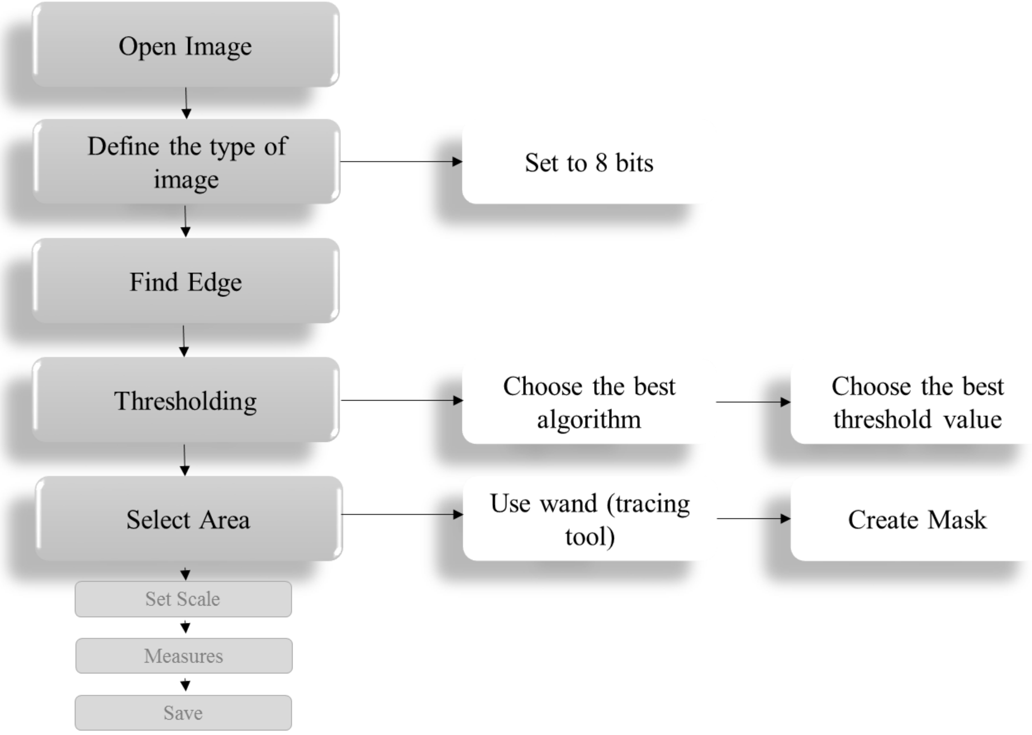


Figure 13: Analysis pipeline in ImageJ, emphasising the pre-processing steps.

The first pre-processing step is setting the colour scale. Despite all the images being saved by the experimentalist as grey colour palette, not all the images were defined as 8-bit colour scale - grey scale. Setting this informs ImageJ that pixel values are between 0 and 255 (in other software it could be from 1 to 256), as set out in the formula:

$$0 \leq P(x, y) \leq 255,$$

where P(x,y) are the pixel coordinates in the image.

Next, it is necessary to find the border of the colony with the maximum precision possible, since there are measurements in ManualC that classify the border of the colony, such as margin, and it is also important to accurately measure the size of the colony. For this task, we used the function “Find Edge”, that emphasis the edge in the image, by sharp changes in intensity in the ROI [57]. This sharp change occurs near to the limit of the background and the object of interest (see Figure 14).

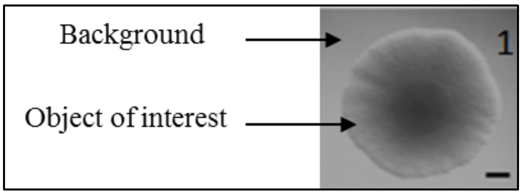


Figure 14: Description of the two regions in an image of the data set.

Find Edge is the first of two operations to separate the ROI from the background. The second operation is thresholding. As described previously, this operation sets the pixels that have lower and upper threshold values, segmenting greyscale images into objects of interest and background. Although this is a widely used operation, it is a complex one. It is necessary to be careful with the thresholding algorithm chosen and also with the threshold value, as it can change the shape of the object of interest [67].

The threshold value is different for each image and the algorithm chosen may also change. When the object of interest identified was not satisfactory, it was therefore necessary to try different methods available in ImageJ. The most frequently used were the Triangle algorithm and the Yen algorithm.

After thresholding it is possible to select only the colony, using the Wand Tool. This tool creates a selection by tracing thresholding objects [57]. In this process, a Mask was created and saved. The mask is an 8-bit image where the pixel value inside the selection is 255 (white) and on outside is 0 (black) [57]. This step is necessary to maintain the selection of the object of interest for another operation in the future, without losing the original image.

Once the region of interest has been selected, the image is ready for the measurement operations.

4.2.2. Extraction of colony automatic features from images

After pre-processing the image, and studying the manually curated features, it was necessary to study the features, here referred to as measurements, which best represent colony morphology. ImageJ with its many available plugins, can provide many measurements to describe an object. The approach used in this work was to try to use the ImageJ measurements that are more similar to those used in manual curation. Therefore we explored ImageJ tools and selected 4 groups of measurements:

- Basic shape;
- Basic pixel value;
- Textural;
- Roughness;

The basic shape measurements seeks to describe numerically the profile of the colonies and is related with ManualC features, Form and Size (see Table 4).

Table 4: Description of basic shape measurements obtained from ImageJ.

Basic Shape Measurements in ImageJ	
Area	Area of selection in square pixels.
Centroid	The center point of the selection. This is the average of the x and y coordinates of all of the pixels in ROI.
Center of Mass	The brightness-weighted average of the x and y coordinates all pixels in the ROI.
Fit ellipse	Fits an ellipse to the ROI. Major and Minor are the primary and secondary axis of the best fitting ellipse. Angle is the angle between the primary axis and a line parallel to the x-axis of the image.
Feret's Diameter	The longest distance between any two points along the ROI boundary, also known as maximum caliper.
Feret Angle	The angle (0-180 degrees) of the Feret's diameter.
MinFeret	The minimum caliper diameter.
Skewness	The third order moment about the mean.
Kurtosis	The fourth order moment about the mean.
Area Fraction	For thresholding images, this is the percentage of pixels in the ROI that have been highlighted in red during the thresholding. For non-threshold images, this is the percentage of non-zero pixels.
Perimeter	The length of the outside boundary of the ROI.
Bounding rectangle	The smallest rectangle enclosing the ROI.
Aspect ratio	The aspect ratio of the particle’s fitted ellipse.
Circularity	Indicate how near ROI is from a perfect circle (1.0).
Roundness	Similar to circularity, largest value 1.0 for a circular object.
Solidity	Whether one object has an irregular border or not.

The basic pixel value, Table 5 is comparable with the ManualC “Surface”. As the name indicates, the basic pixel values describe the values of the pixels in the ROI. Both basic groups of measurements are accessible in the default ImageJ installation, and this was one of the reasons to incorporate these measurements in the automatic analysis.

Table 5: Description of basic pixel value measurements from ImageJ.

Basic Pixel Value Measurements in ImageJ	
Mean	Average grey value within the ROI.
Modal	Most frequently grey value occurring within the ROI. Corresponds to the highest peak in the histogram.
Median	The median value of the pixels in the ROI.
Min	The lower value of the pixel intensity in the ROI.
Max	The upper value of the pixel intensity in the ROI.
Standard Deviation	The standard deviation of the grey values used to generate the mean grey value.
Integrated Density	The sum of the values of the pixels in the ROI.

Textural measurements are metrics designed to quantify the perceived texture, i.e. the spatial arrangement of intensities ROI. The textural metrics evaluated are described in Table 6.Textural measurements result from an Image plugin. It should be noted that the term textural as used in ImageJ and texture from the ManualC are different. In the former, texture is something touchable allied with the surface, and in the latter textural is associated with the variation of the pixel intensity.

Table 6: Examples of textural metrics provided by ImageJ.
Where $p(i,j)$ are the $(i,j)th$ entry in a grey-tone spatial dependence matrix, $p_x(i)$ is the $(i)th$ entry in the marginal-probability matrix obtained by summing the rows of $p(i,j).p_y(j)$ is the $(j)th$ entry in the marginal-probability matrix obtained by summing the column of $p(i,j)$. N_g is the number of distinct grey levels in the quantized image. At last, μ_x, μ_y, σ_x and $\sigma_y$are the means and the standard deviations of p_x and p_y .

Textural		
Angular Second Moment (ASM)	Assessment of homogeneity of the image. High values of ASM occur when the image is very orderly and contains only a few grey levels.	$f_{ASM} = \sum_i \sum_j \left(p(i,j)\right)^2$
Contrast	Quantification of contrast or local intensity variation. Higher the values of the contrast correspond to more variation in the pixel intensity.	$f_{Cont} = \sum_{n=0}^{N_g-1} n^2 \left(\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \right)$ $where \lvert i-j \rvert = n$
Correlation	Measurement of grey level linear dependence between the pixels at the specified positions relative to each other.	$f_{Cor} = \frac{\sum_i \sum_j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
Inverse Difference Moment (IDM)	Distribution of the grey-level in the image. It is high when local grey level is uniform.	$f_{IDM} = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$
Entropy	Entropy indicates the degree of "chaos" in the system. Non-homogeneous scenes have low first order entropy, while homogeneous scenes have high entropy.	$f_E = -\sum_i \sum_j p(i,j) \, log\left(p(i,j)\right)$

Roughness consists of surface irregularities. The input is an image in which the pixel values represent distance, z, to a surface. These irregularities are combined to form a surface texture that can be also quantified [68,69]. In simplified terms, each pixel in the image has 3 values, x, y and the colour intensity, so it is possible to transform a 2D in a 3D graphic where the colour intensity is now represented in the z-axis, and if the process is repeated for all the pixels in an image we obtain a surface in a 3D graphic (Figure 15). After this step, all the variation on this surface is calculated with the roughness calculation plugin. The ImageJ roughness metrics evaluated are described in Table 7.

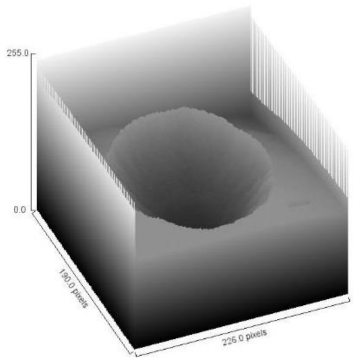


Figure 15: Surface profile plot of image1 (see also Figure 10 and Table 3).
Where z axis represent the colour value for each pixel. A visual plot, that helps to understand the basic notion of surface roughness.

Table 7: Examples of metrics outputted by the surface roughness in ImageJ.
Where $M \times N$ are the rectangular image, $Z(x_k, y_i)$ is the point that represents the vertical distance from the mean line to data point(x_k, y_i).

Roughness Calculation		
Roughness Average	Arithmetic average of absolute values in the image.	$Ra = \frac{1}{MN}\sum_{k=0}^{M-1}\sum_{i=0}^{N-1}\lvert z(x_k, y_i)\rvert$
Root Mean Square	Root mean squared of the pixel values within the image.	$Rq = \sqrt{\frac{1}{MN}\sum_{k=0}^{M-1}\sum_{i=0}^{N-1}[z(x_k, y_i)]^2}$
Skewness	Describes the asymmetry of the height distribution histogram. If R_{sk}=0 a symmetric height distribution of the pixel values is indicated. If R_{sk}<0, it can be a bearing surface with holes and if R_{sk}>0 it can be a flat surface with peaks. Values numerically greater than 1.0 may indicate extreme holes or peaks on the surface.	$Rsk = \frac{1}{MNR_q^3}\sum_{k=0}^{M-1}\sum_{i=0}^{N-1}[z(x_k, y_i)]^3$
Kurtosis	Describes the "peakedness" of the surface topography. It is a descriptor of the shape of the probability distribution.	$Rsu = \frac{1}{MNR_q^4}\sum_{k=0}^{M-1}\sum_{i=0}^{N-1}[z(x_k, y_i)]^4$
Lowest valley	The largest valley depth value of the surface form with the pixel intensity of the image	$Rv = \min_z z(x_k, y_i)$
Highest peak	The largest peak height value. of the surface form with the pixel intensity of the image	$Rp = \max_z z(x_k, y_i)$
Total height	Defined as the height difference between the highest and lowest pixel in the image.	$Rt = Rp - Rv$

All the measurements presented in this sub chapter were used to extract information about the colonies, composing the ImageJ dataset and were used to infer the accuracy of these measurements.

4.3 Data mining

Data mining is the science of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable and predictive models from large-scale data [70]. Image mining can the considered as a sub-section of data mining that deals with the extraction of implicit knowledge, image data relationship, or other patterns not explicitly stored in images. Images have always been used in biological studies. Nowadays, with the capacity to acquire and store biological images, the number of image is growing incessantly. Image mining is one of the great challenges of data mining which connects image processing with data mining [71].

A challenge in image processing is that knowledge cannot be easily transferred from one domain to another, since each image is linked with a specific field. Knowledge designates data classification, clustering or prediction. Some of the bioinformatic's goals are organizing, storing and processing information on molecular and cellular, processes, tissues and organs, individuals, population and society to support the definition of suitable decision making strategies in health care [72–74]. In this thesis, the goal is to extract patterns from experimental data, stored at http://morphocol.org/, using not only the morphology ontology available (ManualC), but also data extracted through image processing. Ultimately, this body of knowledge will improve the existing ontology and thus the information available to design a clinical-decision making system.

4.3.1 Selection of relevant image measurements

After image processing it was necessary to study if all the acquired metrics were relevant to this study. To infer that the data mining software Weka http://www.cs.waikato.ac.nz/ml/weka/ was used [75]. This software provides several methods for identifying those subsets of attributes (measurements) that are predictive of another (target) attribute in the data. The results presented in Table 8 were obtained with a full dataset of images comparing the values of all the features extract using ImageJ.

Table 8: The top 15 ImageJ measurements selected by different algorithms in Weka.
The all measurements extract in ImageJ were submitted to four attribute selection algorithm.

GainInfo	ChiSquared	ReliefF	SVM
Area	Area	Angle	Contrast
Height	Height	FeretAngle	Round
Entropy	Rq	InverseDifferenceMoment	HighestPeak
Perim	Ra	Entropy	Rq
Contrast	Angle	Max	Skew
FeretAngle	Entropy	Circ	Kurt
Rq	IntDen	Mode	LowestValley
Ra	Contrast	XM	Mean
Angle	FeretAngle	YM	YM
Major	Mean	Correlation	Height
Mean	Major	Height	Rsk
StdDev	StdDev	LowestValley	AR
Perim	Perim	Ra	FeretAngle
Kurt	Kurt	Feret	XM
XM	MinFeret	MinFeret	Width

4.3.2 K-means clustering

Clustering is a technique for unsupervised learning, where data with class labels are not available. Clustering is the task of partitioning the data points into natural groups called clusters, such that points within a group are very similar, whereas points across clusters are as dissimilar as possible. Depending on the data and desired cluster characteristics, there are different types of clustering paradigms such as representative-based, hierarchical, density-based, graph-based and spectral clustering [71,72,76].

Here the k-means algorithm was used to perform clustering of the two datasets – ManualC and ImageJ, where k is the number of clusters (groups). Clustering was performed with the data mining software RapidMiner [71,77]. Considering that the datasets contain 111 images, k was set to 17. The maximum number of runs was set to 100 and the max optimization steps taken within each run was set to 1000.

4.3.3 Clustering evaluation metrics

After performing clustering it is necessary to evaluate the quality of the resulting clusters. To do that there are several metrics available in [78]. The metrics used in this work are detailed bellow.

Intra-cluster distance and similarity measurements

The Euclidean distance and the cosine similarity are used to estimate the similarity between instances in the same cluster.

Given two n-dimension instances $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$, the Euclidean distance between them is:

$$d(x_i, x_j) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2},$$

and the average distance for each cluster is given by the arithmetic mean of the distance calculated for every pair of instances in the cluster. In turn, the cosine similarity compares images according to the angle established by their sets of measurements as follows:

$$s(x_i, x_j) = \frac{x_i^T . x_j}{\|x_i\| . \|x_j\|}.$$

The score ranges between [0, 1] such that the more similar the vectors are, the higher the value.

Individual Clustering Quality

The sum of squared error (SSE) computes the combined effect of intra-cluster homogeneity and separation between clusters. Its calculation is as follows:

$$SSE = \sum_{k=1}^{K} \sum_{\forall x_i \in C_k} \|x_i - \mu_k\|^2,$$

where C_k is the set of instances in cluster k, x_i is one instance in the cluster (i.e. an image with a set of measurements), and μ_k is the vector mean of cluster k calculated such that:

$$\mu_{k,j} = \frac{1}{N_k} \sum_{\forall x_i \in C_k} x_{i,j},$$

where $N_k = |C_k|$ is the number of instances belonging to cluster k.

Inter-annotation Clustering Agreement

To quantify the inter-annotation agreement between ImageJ and Manual datasets, we calculate the precision, recall and F-score, common metrics in inter-annotation evaluations, originated in Information Retrieval theory [77].

The precision-recall measure can be used as an external measure for evaluating clusters. The cluster is viewed as the results of a query for a specific class. Precision is the fraction of correctly retrieved instances, while recall is the fraction of correctly retrieved instances out of all matching instances. A combined F-measure can be useful for evaluating a clustering structure [79]. We have adapted the original formulas to our needs.

$$precision = \frac{|\{Image\ ManualC\} \cap \{Image\ ImageJ\ \}|}{|Image\ ManualC|}$$

$$recall = \frac{|\{Image\ ManualC\} \cap \{Image\ ImageJ\ \ \}|}{|Image\ ImageJ\ |}$$

$$F - score = 2 \times \frac{precision \times recall}{precision + recall}$$

Chapter 5. Results and discussion

This chapter is organized as follows: first the filtering of the measurements followed by the description of the data transformation perform on the two datasets, then we compared the datasets and to finalize we present and discussed the clustering results and the inter-annotation agreement.

5.1 Filtering of measurements

From the original dataset of 111 images of colonies, two types of information were extracted and separated in two classes or datasets: (i) ManualC referring to the manual annotation and (ii) ImageJ referring to the measurements obtained through ImageJ (Figure 16). The Results are divided into: (i) evaluation of individual datasets and (ii) comparison of annotations between those datasets. These results are reported according to the overall similarity between images and biological context, notably the species identified, and the comparison with the features most frequently reported in studies of microbial colonies.

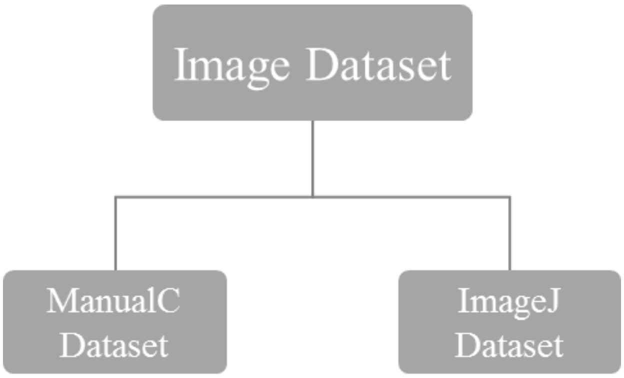


Figure 16: Diagram of the datasets that will be discussed during this chapter.
From the original set of images, two datasets were created depending on how the images were characterized: one was created with the manual annotation features – ManualC; and the other from the measurements obtained from image processing software – ImageJ.

Before any meaningful comparative analysis between ManualC and ImageJ could be made, it was necessary to determine and select measurements extracted from ImageJ that were consistent with the goal of this project. That is, test whether a particular measurement is useful for colony morphology characterization. For this selection, several attribute tests were performed in Weka. The four tests that were ran and the results are available in Table 8 and summarized in Table 9. It is possible to observe that all the groups of measurements are well represented, and indeed some are over-represented. Therefore, a decision was made to exclude measurements that are redundant, for example, those that represent only the position of the ROI in the image. Thus, of Y, X and XM and YM, only XM and YM were kept (classified as important in Weka tests – see (Table 9). Another decision was to exclude measurements such as RawIntDen (sum of the pixel values in the colony), that only brought confusion to the data set. Ultimately, we try to kept only non-redundant, non-ambiguous metrics, for furthers analysis.

Table 9: Measurement most frequently after performing analysis for attribute selection in Weka. It is also present the number of frequency of each measurements

Measurements	Frequencies after Weka tests		
Height	4	YM	2
FeretAngle	4	LowestValley	2
Entropy	3	IntDen	2
Contrast	3	InverseDifferenceMoment	1
Rq	3	Max	1
Ra	3	Circ	1
Angle	3	Mode	1
Mean	3	Correlation	1
Kurt	3	Feret	1
XM	3	Round	1
Perim	2	HighestPeak	1
Area	2	Skew	1
Major	2	AR	1
StdDev	2	Width	1
MinFeret	2	Rsk	1

It was also necessary to have in consideration if the type of image and the type of information that was be extracted from it were consistent. The selection of ImageJ measurements was based on those two assumptions, and with the help of information drawn from the attribute selection tests, 37 measurements were selected to create the ImageJ dataset.

5.2 Data transformation

Due to the different order of magnitude of the measurement's values in the ImageJ dataset (Figure 17), it was necessary to normalize the data. The RapidMiner "Data Transformation" "Normalize" operator was used to normalize the dataset through the Z- transformation method. This transformation increasing the accuracy of the clustering. An important data transformation step was also done to the ManualC dataset. As previously mentioned the manualC features are nominal and needed to be transformed into numerical values. Thus, these nominal values were transformed into binary data using RapidMiner's "Data Transformation" "Nominal to numerical" operator.

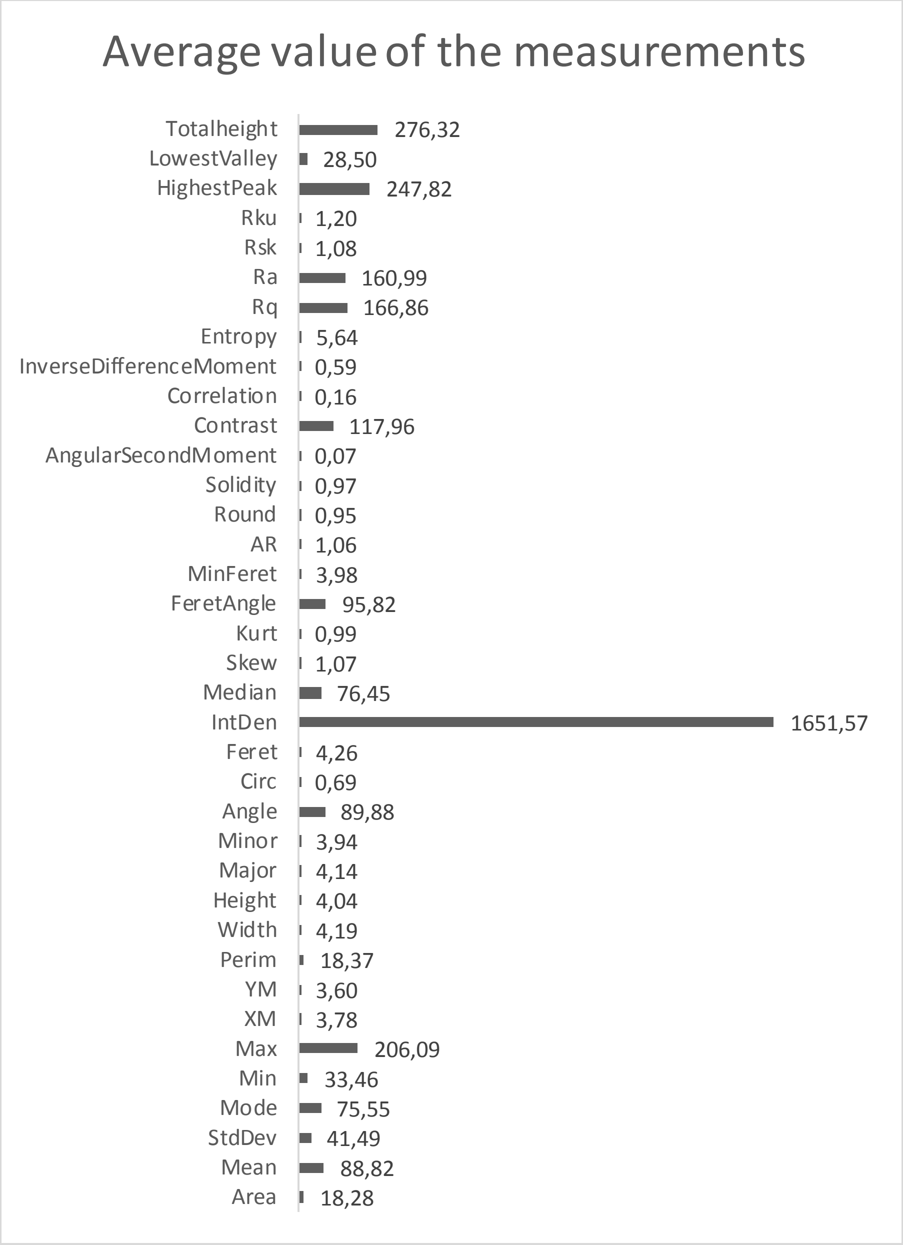


Figure 17: Average value of each ImageJ measurement.
It is visible the range of orders of magnitude between the measurements.

5.3 Visual comparison of datasets

Before proceeding to cluster the data, a preliminary empirical analysis was done to have some assurances about the validity of our comparisons. With that goal, the following questions were posed: is there any measure from ImageJ dataset that can describe a feature from ManualC dataset? Is there any relation between ImageJ dataset measurements and ManualC dataset features? Is there any relationship between this measurements/feature and the species? Following this line of thoughts we tried to find out if for each of the outliers of the ManualC features there was a corresponding value in the ImageJ measurements. The outliers were chosen because there are fewer images for each category, and with more striking features, and thus it was easier to make a visual comparison. This analysis was not performed for to the feature “elevation” due to the lack of variability: all the images are classified as “flat”. Similarly, for the feature “Opacity”, there was only one image with the classification “transparent” whereas the rest of the images are “opaque”, and thus this feature was not analysed.

For the feature “form”, the classification of “irregular” was considered an outlier, because only tree images classified as such. The colonies in these images are formed by the bacteria *Pseudomonas aeruginosa*. This manual classification relates to the ImageJ measurements “cir”, “ar”, “round”, “solidity”, for example. In the case of “ar” (aspect ratio – see Table 4), the “irregular” images have the highest values for this measurement, and the smallest “round” values. In other words, the colonies classified as “irregular” are the ones with worst circularity (see definition in Table 4). The values for “solidity”, a measurement that classifies whether an object is irregular or not, also confirmed that the manualC and the ImageJ classification are consistent.

The feature “sheath” has six different classifications: “entire” (61 images), “undulate” (42 images), “lobulated” (tree images), “erose” (two images), “irregular” (two images) and “curled” (one image). Comparing species vs classification, is possible to deduce that all the *S. aureus*, *E. coli* and *D. pigrum* are “entire”. The only variability for this feature is observed in *P. Aeruginosa* colonies. Focusing on these colonies, and the ImageJ measurements, for “Perim” the “undulate” colonies had the largest values, and the smallest for “circ”. This indicates that these are the colonies with the least perfect circle (see Table 4). In term of “Feret” this images had the biggest values.

The next analysis is dedicated to the feature “Surface”. “Heterogeneous” is an outlier classification of this manualC feature with 17 colonies classified as such. All these colonies are *P. aeruginosa*, but since there are in total 68 *P. aeruginosa* colonies, it is very unlikely that there is a correlation between species and “Surface”. This feature is linked to pixel value measurements (Table 5) but it did not correlate with the values in our ImageJ dataset. However, when comparing “Surface” with entropy, it was notable that all the “heterogeneous” images had entropy values larger than the average (5,64). Also, the “InverseDifferenceMoment” was inferior to the average value of the measurements of the "heterogeneous" images. The above observations lead us to speculate whether there is a relationship between the feature "Surface" and textural measurements.

From the beginning it was assumed that "texture" was the ImageJ analogue to textural measurements (Table 6). The feature "texture" also had outliers for "rough and wrinkled" (eight images), "smooth and wrinkled" (seven images) and "wrinkled" (two images). In terms of textural measurements, for instance, the "correlation" values for the "smooth and wrinkled" colonies were superior to the observed average value (0,16). In the case of "InverseDifferenceMoment" all the "rough and wrinkled" and "smooth and wrinkled" colonies have values inferior to the average (0,59). Also for this measurement, the "smooth" colonies had usually the largest values. The same pattern was observed for "entropy", for which all of the outliers had higher values than the average. All the colonies from *S. aureus* and *D. pigrum* were classified as having a "smooth" texture. Two of the tree images of *E.coli* colonies were classed as "rough" and once again the largest variability in this feature comes from *P. aeruginosa colonies.*

The feature "sheath" is one of the most complicated to relate to the manually extracted measurements. Margin is described as being "a morphological quality inhering in a colony by virtue of having a closely enveloping part or structure after the margin and around the colony"[80]. Since it is unlikely to have measurements in our ImageJ dataset, we ignored the existence of a relationship between "sheath" from ManualC with any ImageJ measurement. When relating species with "sheath", it was possible to infer that for all the *S. aureus*, *E. coli* and *D. pigrum* this feature was noted as "absence" and therefore the variability was observed only for *P. aeruginosa*.

In terms of the feature "consistency", there are only tree colonies classified as "mucoid" and all from *P. aeruginosa*. There were *à priori* no measurements related with this feature, so to this analysis is necessary to go over all the measurements in the ImageJ dataset. For the measurements "area", "stdDev", "max","Perim", "width", "heigth", "Major", "Minor", "Feret", "contrast", "entropy", "Rsk", "Rku" and "HighestPeak" this type of colonies have the largest values and the lowest values of "mode", "min", "cir" and "median. The "Correlation" values, are very similar between them.

For the feature "size", the images are well distributed: 59 are "small", and 52 are "large". All the *S. aureus* and *E. coli* are "large" and all of the *D. pigrum* small and once again the variability occurs for the the *P. aeruginosa colonies.* In general this feature is well correlated with the measurements: "Area", "Perim" and "Feret".

5.4 Clustering

Evaluation of the datasets is based on k-means clustering, which partitions the instances (i.e. images with features) into a pre-defined number of clusters (k), reaching for the best compromise between intra-cluster homogeneity and inter-cluster dissimilarity [81–83].

Clustering results are presented in Table 10 showing the number of images per cluster, the internal homogeneity of the clusters (Euclidean distance and cosine similarity) and the sum of squared error (SSE), i.e. overall performance. Then, the inter-annotation agreement between ImageJ and manually observed features was assessed.

Table 10: Results of k-means partitioning for the ImageJ dataset.
Number of images per cluster and the SSE of each algorithm used. Cluster identifiers are generated by RapidMiner and have no correspondence across similarity metrics.

	ImageJ		ManualC	
	Euclidean	Cosine	Euclidean	Cosine
Cluster 1	8	11	33	33
Cluster 2	1	1	5	5
Cluster 3	2	4	3	3
Cluster 4	4	9	1	1
Cluster 5	14	2	1	1
Cluster 6	1	6	6	6
Cluster 7	18	4	5	5
Cluster 8	3	13	5	5
Cluster 9	6	14	4	4
Cluster 10	6	5	3	3
Cluster 11	8	9	5	5
Cluster 12	8	8	17	17
Cluster 13	18	3	14	14
Cluster 14	1	6	1	1
Cluster 15	4	6	5	5
Cluster 16	1	3	1	1
Cluster 17	8	7	2	2
SSE	0,099	0,077	0,144	0,144

The results of the k-means clustering are presented in Table 10. SSE values are fairly similar, the "average within distance" for cosine similarity (-32.582) is nearer to zero than the Euclidian distance (-42.276), and in terms of the Davies-Bouldin index the clustering performed with Euclidian distance is equal to -1,046 and -1.195 for cosine similarity clustering. We repeated the same exercise to the ManualC dataset. Interestingly, the results are the same despite the different algorithm chosen to perform the k-means clustering. On both tests, the Davies-Bouldin index equals to -0.708 and the avg. within cluster distance is -5.409. In Table 11 we present the clustering results per images.

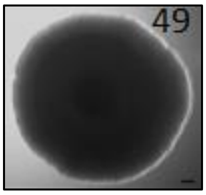
Table 11: All the clustering results performed on the 2 datasets, with 2 different k-means algorithms. Where K=17 and ID is the images identification number.

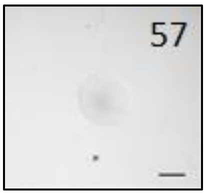
ID	Euclidean_ManualC	Euclidean_ImageJ	Cosine_ManualC	Cosine_ImageJ
1	cluster_13	cluster_13	cluster_13	cluster_16
2	cluster_13	cluster_15	cluster_13	cluster_16
3	cluster_4	cluster_4	cluster_4	cluster_6
4	cluster_5	cluster_4	cluster_5	cluster_6
5	cluster_8	cluster_4	cluster_8	cluster_6
6	cluster_9	cluster_7	cluster_9	cluster_1
7	cluster_9	cluster_5	cluster_9	cluster_9
8	cluster_9	cluster_7	cluster_9	cluster_1
9	cluster_10	cluster_17	cluster_10	cluster_10
10	cluster_10	cluster_15	cluster_10	cluster_14
11	cluster_1	cluster_5	cluster_1	cluster_9
12	cluster_1	cluster_7	cluster_1	cluster_17
13	cluster_1	cluster_7	cluster_1	cluster_1
14	cluster_1	cluster_5	cluster_1	cluster_9
15	cluster_1	cluster_7	cluster_1	cluster_17
16	cluster_1	cluster_7	cluster_1	cluster_1
17	cluster_1	cluster_5	cluster_1	cluster_9
18	cluster_1	cluster_7	cluster_1	cluster_17
19	cluster_1	cluster_7	cluster_1	cluster_1
20	cluster_1	cluster_5	cluster_1	cluster_9
21	cluster_1	cluster_7	cluster_1	cluster_17
22	cluster_1	cluster_7	cluster_1	cluster_1
23	cluster_1	cluster_5	cluster_1	cluster_9
24	cluster_1	cluster_5	cluster_1	cluster_9
25	cluster_1	cluster_7	cluster_1	cluster_1
26	cluster_1	cluster_5	cluster_1	cluster_9
27	cluster_1	cluster_7	cluster_1	cluster_17
28	cluster_1	cluster_7	cluster_1	cluster_1
29	cluster_1	cluster_5	cluster_1	cluster_9
30	cluster_1	cluster_7	cluster_1	cluster_17
31	cluster_1	cluster_7	cluster_1	cluster_1
32	cluster_1	cluster_5	cluster_1	cluster_9
33	cluster_1	cluster_5	cluster_1	cluster_9
34	cluster_1	cluster_7	cluster_1	cluster_1
35	cluster_1	cluster_5	cluster_1	cluster_9
36	cluster_1	cluster_7	cluster_1	cluster_17
37	cluster_1	cluster_7	cluster_1	cluster_1
38	cluster_11	cluster_17	cluster_11	cluster_3
39	cluster_1	cluster_13	cluster_1	cluster_11
40	cluster_12	cluster_13	cluster_12	cluster_11
41	cluster_9	cluster_17	cluster_9	cluster_10
42	cluster_12	cluster_13	cluster_12	cluster_11
43	cluster_12	cluster_13	cluster_12	cluster_11

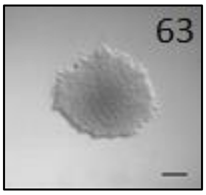
ID	Euclidean_ManualC	Euclidean_ImageJ	Cosine_ManualC	Cosine_ImageJ
44	cluster_11	cluster_17	cluster_11	cluster_10
45	cluster_13	cluster_13	cluster_13	cluster_16
46	cluster_12	cluster_13	cluster_12	cluster_11
47	cluster_11	cluster_17	cluster_11	cluster_10
48	cluster_13	cluster_10	cluster_13	cluster_4
49	cluster_12	cluster_16	cluster_12	cluster_4
50	cluster_11	cluster_17	cluster_11	cluster_10
51	cluster_13	cluster_13	cluster_13	cluster_11
52	cluster_14	cluster_13	cluster_14	cluster_11
53	cluster_11	cluster_17	cluster_11	cluster_3
54	cluster_13	cluster_13	cluster_13	cluster_11
55	cluster_12	cluster_13	cluster_12	cluster_11
56	cluster_13	cluster_13	cluster_13	cluster_14
57	cluster_15	cluster_2	cluster_15	cluster_7
58	cluster_6	cluster_13	cluster_6	cluster_14
59	cluster_6	cluster_13	cluster_6	cluster_14
60	cluster_8	cluster_13	cluster_8	cluster_7
61	cluster_6	cluster_1	cluster_6	cluster_14
62	cluster_15	cluster_13	cluster_15	cluster_7
63	cluster_13	cluster_6	cluster_13	cluster_3
64	cluster_16	cluster_5	cluster_16	cluster_9
65	cluster_15	cluster_5	cluster_15	cluster_9
66	cluster_8	cluster_17	cluster_8	cluster_3
67	cluster_13	cluster_13	cluster_13	cluster_14
68	cluster_13	cluster_10	cluster_13	cluster_4
69	cluster_13	cluster_13	cluster_13	cluster_7
70	cluster_2	cluster_10	cluster_2	cluster_4
71	cluster_12	cluster_15	cluster_12	cluster_12
72	cluster_12	cluster_11	cluster_12	cluster_12
73	cluster_12	cluster_11	cluster_12	cluster_12
74	cluster_12	cluster_11	cluster_12	cluster_12
75	cluster_12	cluster_15	cluster_12	cluster_6
76	cluster_12	cluster_4	cluster_12	cluster_6
77	cluster_12	cluster_11	cluster_12	cluster_12
78	cluster_12	cluster_11	cluster_12	cluster_12
79	cluster_12	cluster_11	cluster_12	cluster_6
80	cluster_12	cluster_11	cluster_12	cluster_12
81	cluster_12	cluster_11	cluster_12	cluster_12
82	cluster_2	cluster_9	cluster_2	cluster_8
83	cluster_3	cluster_3	cluster_3	cluster_4
84	cluster_7	cluster_9	cluster_7	cluster_8
85	cluster_6	cluster_9	cluster_6	cluster_8
86	cluster_7	cluster_10	cluster_7	cluster_4
87	cluster_2	cluster_9	cluster_2	cluster_8
88	cluster_8	cluster_12	cluster_8	cluster_15

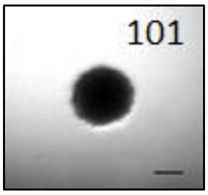
ID	Euclidean_ManualC	Euclidean_ImageJ	Cosine_ManualC	Cosine_ImageJ
89	cluster_2	cluster_1	cluster_2	cluster_8
90	cluster_3	cluster_10	cluster_3	cluster_4
91	cluster_7	cluster_1	cluster_7	cluster_8
92	cluster_13	cluster_1	cluster_13	cluster_8
93	cluster_3	cluster_3	cluster_3	cluster_4
94	cluster_1	cluster_12	cluster_1	cluster_15
95	cluster_8	cluster_12	cluster_8	cluster_5
96	cluster_6	cluster_9	cluster_6	cluster_8
97	cluster_6	cluster_1	cluster_6	cluster_8
98	cluster_1	cluster_12	cluster_1	cluster_15
99	cluster_1	cluster_12	cluster_1	cluster_15
100	cluster_15	cluster_8	cluster_15	cluster_13
101	cluster_10	cluster_14	cluster_10	cluster_2
102	cluster_13	cluster_1	cluster_13	cluster_8
103	cluster_1	cluster_12	cluster_1	cluster_15
104	cluster_1	cluster_12	cluster_1	cluster_5
105	cluster_17	cluster_8	cluster_17	cluster_13
106	cluster_2	cluster_1	cluster_2	cluster_8
107	cluster_15	cluster_12	cluster_15	cluster_15
108	cluster_17	cluster_8	cluster_17	cluster_13
109	cluster_7	cluster_10	cluster_7	cluster_4
110	cluster_13	cluster_9	cluster_13	cluster_8
111	cluster_7	cluster_1	cluster_7	cluster_8

The algorithm of k-means aims to achieve the best performance by committing to a trade-off between intra-cluster homogeneity and inter-cluster separability. As such, smaller clusters, namely one or two image size clusters, should identify images that detach from the common morphology represented in the datasets. In this case, it is interesting to inspect the images that are isolated in clusters 2, 6, 14 and 16, using the Euclidean distance present in **Figure 18**.









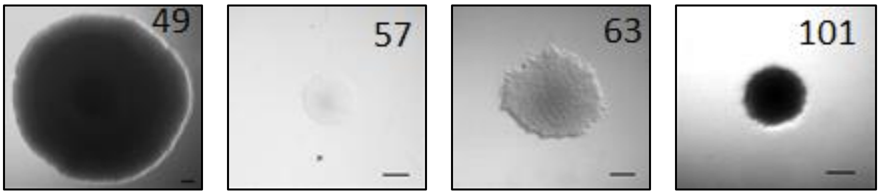


Figure 18: Images clustered in single-cluster by Euclidean K-means. These images belong to the clusters 2, 6, 14 and 16 respectively.

To try to understand the reason why these images were isolated in their own clusters, the ImageJ dataset values were inspected. Image 49 was by far, the largest image in the dataset, with an “area” equivalent to 143,99. Image 57 is extremely bright, i. e., near to white. The selection of the ROI of image 57 was made by manual selection, since automatic selection of ROIs was disabled. Perhaps enabling it would have avoided this ROI. Due to the high brightness of the image, the measurements for colony form, such as “cir” and “solidity”, are not accurate and cannot be used. We can speculate that the high intensity of the pixels is one of the reasons for the image to appear in a unitary cluster. This assumption was confirmed after analysing all data from the ImageJ dataset, and Image 57 has, for example, the lowest “Stddev” value in the dataset, the largest “Mean” and “Min. Similarly, Image 63 appears to be isolated because it had higher values of pixel intensity such as “HighestPeak” (193) or “Stddev” (24,89). In other words, like image 57, image 63 is a bright image.

Image 101 is the only image that appears in a unitary cluster when clustering is performed with cosine similarity. It had the largest “Stddev” value in the dataset, the lower “min” and the upper “max”. These 3 measurements indicate that it is the image with the largest colour spectrum in the dataset. This image also had a lower “solitidy”, “cir”, “AR” and the biggest “Round”, indicating that the colony had an irregular border. Table 12 is a summary of the clustering results regarding only the ImageJ dataset. In general, and despite the different clusters ids, the images clustered similarly. The images in the Euclidean cluster 5 are the same of the Cosine cluster 9. Cluster 8 and cluster 13, Euclidian and Cosine respectively, also share the same images. Images in that appear conjointly in one Euclidean cluster can also appear as separated Cosine clusters. For example, the images in Euclidean cluster 7, are separated in Cosine cluster 1 and cluster 17. The only Euclidean cluster where the images were scattered in several Cosine clusters, is the cluster 15 (Figure 19).

Table 12: Euclidean distance clustering *versus* Cosine Similarity clustering.
The data shows the number of images that were similarly clustered in the two clustering methods.

Euclidean distance clusters	Images	Cosine Similarity clusters	Images	Number of images in common
C1	89 ; 106 ; 92 ; 102 ; 61 ; 97 ; 91 ; 111	C8	82 ; 84 ; 85 ; 87 ; 89 ; 91 ; 92 ; 96 ; 97 ; 102 ; 106 ; 110 ; 111	7
C3	83 ; 93	C4	48 ; 49 ; 68 ; 70 ; 83 ; 86 ; 90 ; 93 ; 109	2
C5	7 ; 11 ; 14 ; 17 ; 20 ; 23 ; 24 ; 26 ; 29 ; 32 ; 33 ; 35 ; 64 ; 65	C9	7 ; 11 ; 14 ; 17 ; 20 ; 23 ; 24 ; 26 ; 29 ; 32 ; 33 ; 35 ; 64 ; 65	14
C7	12 ; 13 ; 15 ; 16 ; 18 ; 19 ; 21 ; 22 ; 25 ; 27 ; 28 ; 30 ; 31 ; 34 ; 36 ; 37 ; 6 ; 8	C1	6 ; 8 ; 13 ; 16 ; 19 ; 22 ; 25 ; 28 ; 31 ; 34 ; 37	11
		C17	12 ; 15 ; 18 ; 21 ; 27 ; 30 ; 36	7
C8	100 ; 105 ; 108	C13	100 ; 105 ; 108	3
C9	82 ; 87 ; 110 ; 85 ; 96 ; 84	C8	82 ; 84 ; 85 ; 87 ; 89 ; 91 ; 92 ; 96 ; 97 ; 102 ; 106 ; 110 ; 111	6
C10	70 ; 48 ; 68 ; 90 ; 86 ; 109	C4	48 ; 49 ; 68 ; 70 ; 83 ; 86 ; 90 ; 93 ; 109	6
C11	72 ; 73 ; 74 ; 77 ; 78 ; 79 ; 80 ; 81	C12	71 ; 72 ; 73 ; 74 ; 77 ; 78 ; 80 ; 81	7
C12	94 ; 98 ; 99 ; 103 ; 104 ; 107 ; 88 ; 95	C15	88 ; 94 ; 98 ; 99 ; 103 ; 107	6
C13	39 ; 40 ; 42 ; 43 ; 46 ; 55 ; 1 ; 45 ; 51 ; 54 ; 56 ; 67 ; 69 ; 52 ; 62 ; 58 ; 59 ; 60	C11	39 ; 40 ; 42 ; 43 ; 46 ; 51 ; 52 ; 54 ; 55	9
		C14	88 ; 94 ; 98 ; 99 ; 103 ; 107	4
		C7	57 ; 60 ; 62 ; 69	3
		C16	1 ; 2 ; 45	2
C17	38 ; 44 ; 47 ; 50 ; 53 ; 66 ; 41 ; 9	C3	38 ; 53 ; 63 ; 66	3
		C10	9 ; 41 ; 44 ; 47 ; 50	5

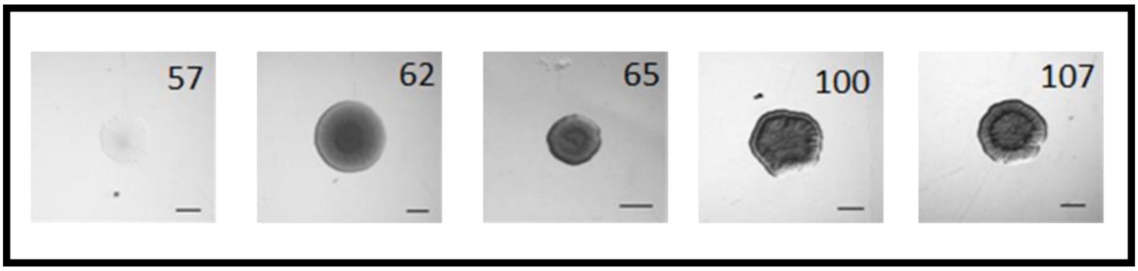


Figure 19: Images in Euclidian Distance cluster C15 that appear in separate Cosine Similarity clusters.

In turn, large size clusters could be explained by a high similarity between the images or, for some of the images, by a reasonable similarity that does not compromise overall performance. The 14 images clustered in Euclidean cluster 5 are the same as in the Cosine cluster 9. Another big

Euclidean cluster is the cluster C7 that in Cosine clustering are separated in two clusters – cluster 1 and cluster 17. The biggest group in Euclidian clustering is cluster 13 with 18 images, but in Cosine clustering this cluster appears fragmented in four different clusters (cluster 7, cluster11, cluster 14 and cluster16).

Regarding the species in each cluster, in Euclidean cluster 7 for example, all the images are from *D. pigrum*. The other *D. pigrum* colonies appear in Euclidean cluster 5, along with 2 *P. aeruginosa* images (i64 and i65). In the biggest Euclidean cluster, cluster 13, all images are from *P. aeruginosa* bar one from *E. Coli*. Performing the same analysis to the "big" Cosine Clusters, in cluster 1 all the images are from *D. pigrum* colonies; all the images in cluster 8 and cluster 11 are *P. aeruginosa*. From these observations, it could be inferred that cosine clustering is better suited to separated images in terms of species.

5.5 Inter-annotation agreement

After the initial inspection of the clustering methods, the automatic quantification and manual annotation of colonies were compared. We speculated that the level of agreement between the two sets of clusters, i.e. the number of images that are grouped together, is affected by the nature of the features/measurements. The measurements extracted from ImageJ are context-blind (i.e. do not refer to any specifics on microbial colonies) and highly dependent on the image quality and focus (i.e. image acquisition and pre-processing). In turn, manual annotations are higher level descriptions of morphological features that researchers are able to visualize and consider relevant to describe microbial communities. Therefore, manual and automatic features are not directly comparable. The inter-annotation agreement evaluation refers to manual annotations as "the eye of the beholder", i.e. cluster distribution is inspected in terms of the features manually observed. It is interesting to investigate whether the fine grained and context independent analysis provided by computational tools brings forward more discriminating ability, or calls attention to non-visually observable or unreported aspects of the colonies.

In the Table 13 there is a comparison between the clustering (Euclidean Distance) performed with ManualC dataset and the clustering performed with ImageJ dataset, assuming that the ManualC clustering is the ground truth. The table also contains the values of clustering agreement between the annotations, for each cluster. In Figure 20 there is a visualization of the distribution of images in the clusters of the two datasets.

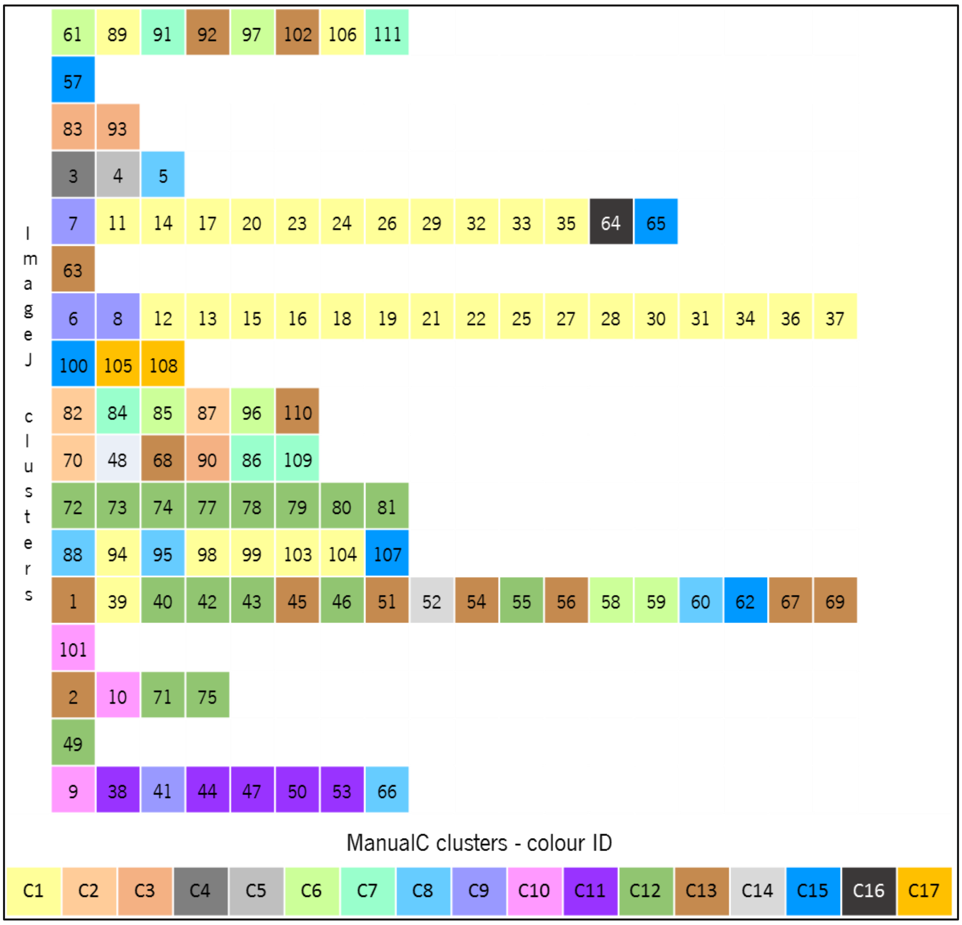
Table 13: Assessment of Euclidian distance clustering quality representing the level of agreement between the annotations
All of presented the inter-annotation agreement metrics described in 4.3.3.

Manual curation	Number of images	ImageJ cluster with more images in common	Number of image in common	Precision	Recall	F-measure
Cluster 1	33	C7	16	0,485	0,889	0,627
Cluster 2	5	C9	2	0,400	0,333	0,364
Cluster 3	3	C3	2	0,667	1,000	0,800
Cluster 4	1	C4	1	1,000	0,250	0,400
Cluster 5	1	C4	1	1,000	0,250	0,400
Cluster 6	6	C9	2	0,333	0,333	0,333
Cluster 7	5	C10	2	0,400	0,333	0,364
Cluster 8	5	C12	2	0,400	0,250	0,308
Cluster 9	4	C7	2	0,500	0,111	0,182
Cluster 10	3	C14	1	0,333	1,000	0,500
Cluster 11	5	C17	5	1,000	0,625	0,769
Cluster 12	17	C11	8	0,471	1,000	0,640
Cluster 13	14	C13	7	0,500	0,389	0,438
Cluster 14	1	C13	1	1,000	0,056	0,105
Cluster 15	5	C2	1	0,200	1,000	0,333
Cluster 16	1	C5	1	1,000	0,071	0,133
Cluster 17	2	C8	2	1,000	0,667	0,800

An example of good agreement between the clusters of the datasets, are the ManualC cluster 11 and the ImageJ cluster 17. All the images in the first cluster are in the second cluster, although cluster 17 has 3 more images. In terms of species composition, all the images involved are from *P. aeruginosa*, but because this is the most frequent species in the dataset (68 in 111), no conclusions can be made. An opposite example is the ManualC cluster 15, in which all its images are dispersed in different ImageJ clusters (Figure 20).

It is also important to note the ManualC cluster 1 - the largest cluster. Despite the composing images being dispersed throughout several ImageJ clusters, 16 of its images are in one cluster (cluster 7) and eleven in other cluster (cluster 5). In addition, in ManualC cluster 1 there are two species, *D. pigrum* and *P. Aeruginosa*. It should also be added that, in ImageJ clusters, all the *D. pigrum* are in cluster 5 and cluster 7, whereas *P. aeruginosa* is disperse in other clusters. Also, all the images in ImageJ cluster 5 and cluster 7 have the same manualC features "circular; entire;

homogeneous; smooth; absence; opaque; flat; dry; small" and are all from *D. Pigrum*. Nonetheless, not all the images with these features are *D. pigrum.*



61 89 91 92 97 102 106 111
57
83 93
3 4 5
7 11 14 17 20 23 24 26 29 32 33 35 64 65
63
6 8 12 13 15 16 18 19 21 22 25 27 28 30 31 34 36 37
100 105 108
82 84 85 87 96 110
70 48 68 90 86 109
72 73 74 77 78 79 80 81
88 94 95 98 99 103 104 107
1 39 40 42 43 45 46 51 52 54 55 56 58 59 60 62 67 69
101
2 10 71 75
49
9 38 41 44 47 50 53 66

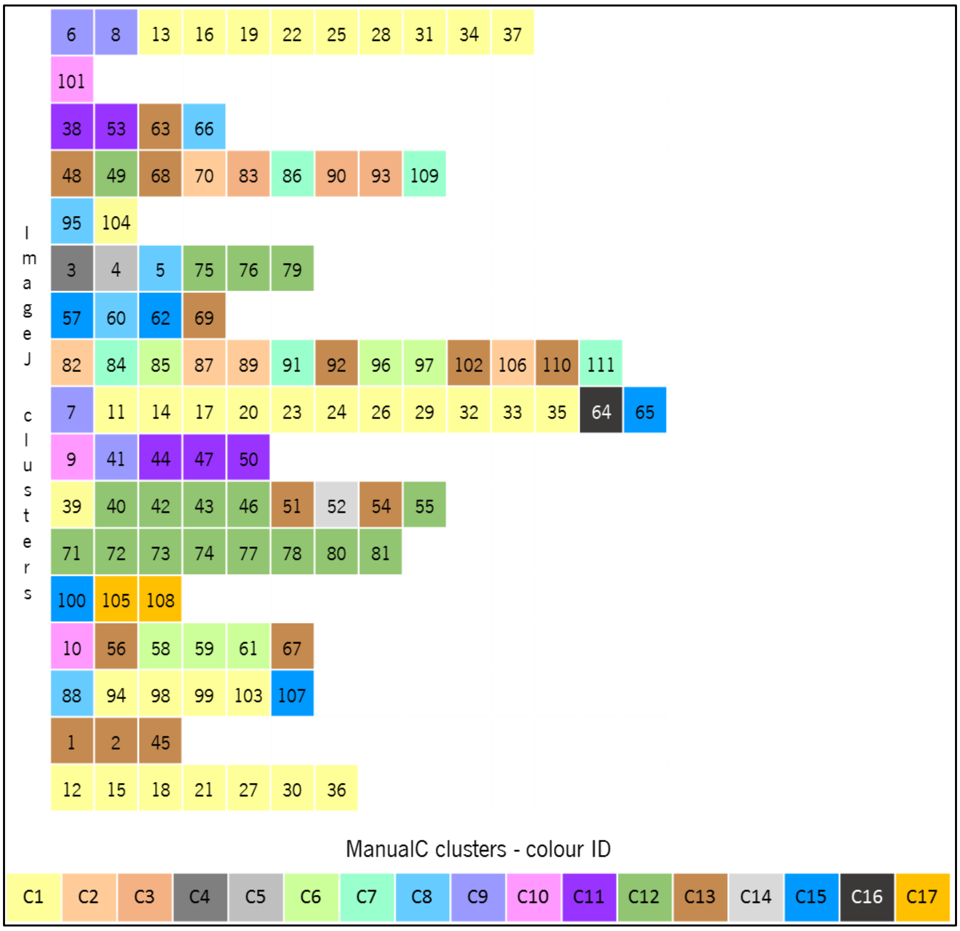
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14 C15 C16 C17

Figure 20: Distribution of images in ImageJ and manualC clusters generated with Euclidian distance. The row on the bottom is a colour code of the ManualC clusters. The other rows each represent an ImageJ cluster and the distribution of images in that cluster. Clustering has performed with Euclidean distance.

The comparison between manual and ImageJ was also performed for Cosine Similarity clustering. The results are presented in Table 14, again considering the ManualC clustering as the ground truth. The values of similarity between ManualC and ImageJ clusters is also present. Figure 21 is a visual representation of the distribution of images in the clusters generated for the two annotations.

Table 14: Assessment of Cosine Similarity clustering quality representing the level of agreement between the annotations

Manual curation	Number of images	ImageJ cluster with more images in common	Number of image in common	Precision	Recall	F-measure
Cluster 1	33	C9	11	0,333	0,786	0,468
Cluster 2	5	C8	4	0,800	0,308	0,444
Cluster 3	3	C4	3	1,000	0,333	0,500
Cluster 4	1	C6	1	1,000	0,167	0,286
Cluster 5	1	C6	1	1,000	0,167	0,286
Cluster 6	6	C14	3	0,500	0,500	0,500
Cluster 7	5	C6	3	0,600	0,500	0,545
Cluster 8	5	C5	1	0,200	0,500	0,286
Cluster 9	4	C1	2	0,500	0,182	0,267
Cluster 10	3	C2	1	0,333	1,000	0,500
Cluster 11	5	C10	3	0,600	0,600	0,600
Cluster 12	17	C12	8	0,471	1,000	0,640
Cluster 13	14	C8	3	0,214	1,000	0,353
Cluster 14	1	C11	1	1,000	0,111	0,200
Cluster 15	5	C7	2	0,400	0,500	0,444
Cluster 16	1	C9	1	1,000	0,071	0,133
Cluster 17	2	C13	2	1,000	0,667	0,800



6 8 13 16 19 22 25 28 31 34 37
101
38 53 63 66
48 49 68 70 83 86 90 93 109
95 104
3 4 5 75 76 79
57 60 62 69
82 84 85 87 89 91 92 96 97 102 106 110 111
7 11 14 17 20 23 24 26 29 32 33 35 64 65
9 41 44 47 50
39 40 42 43 46 51 52 54 55
71 72 73 74 77 78 80 81
100 105 108
10 56 58 59 61 67
88 94 98 99 103 107
1 2 45
12 15 18 21 27 30 36

Figure 21: Distribution of images in ImageJ and manualC clusters generated with Cosine similarity. The row on the bottom is a colour code of the ManualC clusters. The other rows each represent an ImageJ cluster and the distribution of images in that cluster. Clustering has performed with Euclidean distance.

The manualC clustering categorised four clusters with a single image – cluster 4, cluster 5, cluster 14 and cluster 16. Careful inspection of the ManualC data, reveals that these four images have unique combinations of features (Figure 22), providing some explanation for their clustering. Curiously, they are all P. aeruginosa colonies. Also when clustering is performed for the ImageJ dataset, these images were grouped with other images, see Table 14, and analysing the ImageJ data, their features are consistent with those of the images with which they are clustered together.

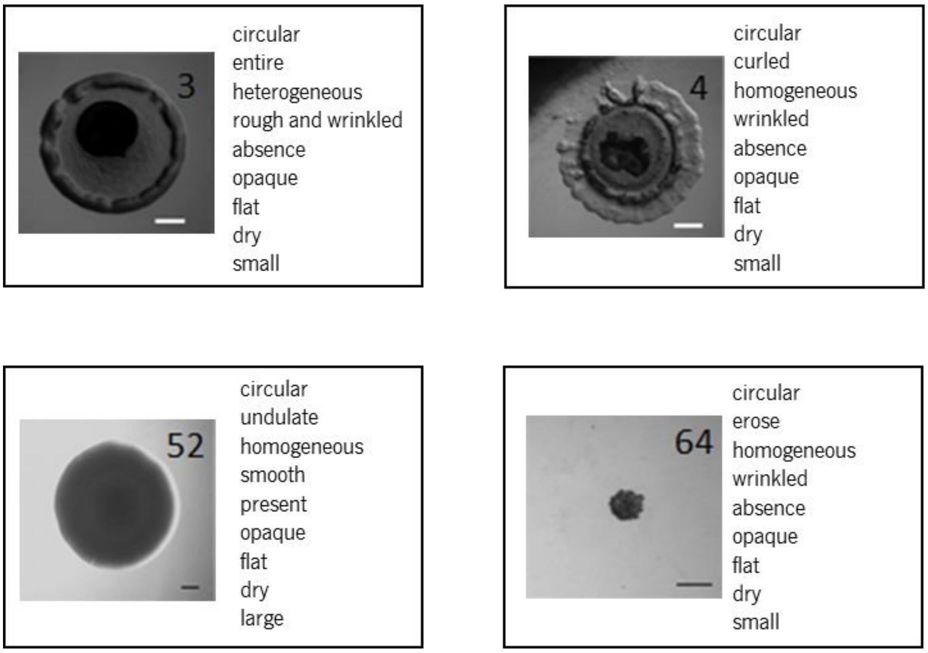


Figure 22: Images in unitary ManualC clusters (Cosine similarity).

Image 3 is in the cluster 4, Image 4 in cluster 5, image 52 in cluster 14 and image 64 in cluster 16

The clusters that are more consistent between ManualC and ImageJ are, respectively, cluster 17 and cluster 13, where the two images in cluster 17 are grouped with other in cluster 13 (Figure 23). These are all images from *P. aeruginosa*. In terms of ImageJ data, the tree images have the same range of values for shape measurements, pixel intensity and size.

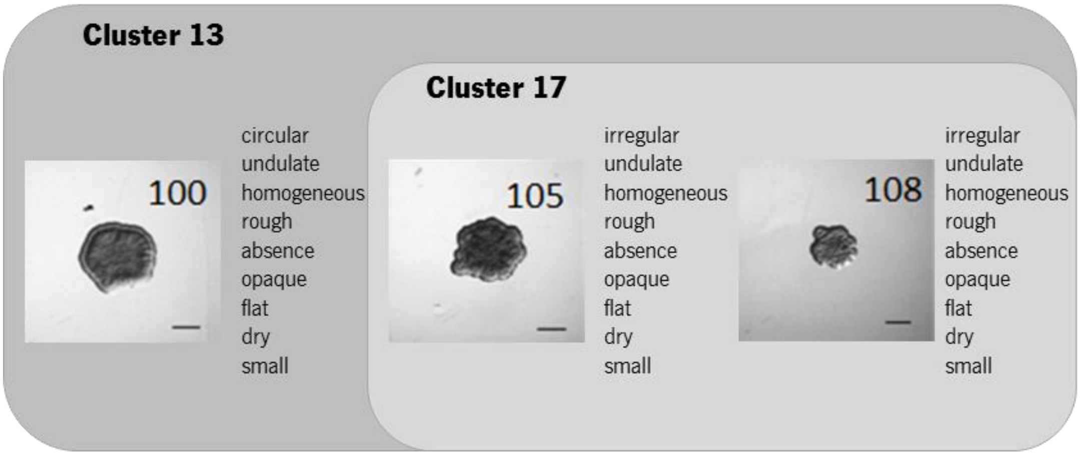


Figure 23: An example of good inter-cluster agreement between ManualC and ImageJ.
The images in cluster 17 (ManualC dataset) are in cluster 13 when the clustering is performed for the ImageJ dataset. Image 100 only differs from the others in the “form” feature.

Chapter 6. Conclusion and Future work

All the aspects presented and discussed in this work are important for assessing whether automatic tools can add or substitute manual curation in the process of morphotyping. Moreover, this study explores the ability of image annotations to support predictions, namely for clinical decision-making.

After studying the software tools, it is possible to conclude that despite the lack of specific software available for this type of colony morphology analysis, ImageJ is a good solution due to the extensibility provided by plugins. In case of colony counting, Cellprofiler, was the software tested that gave better results despite the difficulties in configuring the pipeline. In addition, it is possible to count the colonies by size or by pixel intensity, for example. In the case of biofilm structure analysis, there are several software options, and the choice depends on the "wet lab" protocol followed. But the majority of software tools studied have difficulties with shadows or outliers. If a software does not feature pre-processing tools to deal with this problem, one possible solution is to combine different software tools.

Clustering of the two data sets of measurements was performed with two different methods: Cosine clustering and Euclidian Distance. Our analysis indicates that the clusters obtained for each did not differ significantly, but Cosine similarity should be used in future experiments to reduce entropy.

In this work, it is evident that neither the clusters obtained for the manual curation features or the chosen set of ImageJ measurements, were able to segregate clusters with single species. It was notable that *P. aeruginosa* colonies are the ones with more morphological diversity. Some of the ImageJ measurements can describe well particular manual features, but in the future ImageJ measurements that are not necessarily related to the manual annotations should also be considered. Some available ImageJ plugins that could be explored are for example:

- "Fractal Dimension and Lacunarity", http://rsbweb.nih.gov/ij/plugins/fraclac/fraclac.html,
- "Granulometry",http://rsbweb.nih.gov/ij/plugins/granulometry.html,
- "Fractal Surface Measurement", http://rsbweb.nih.gov/ij/plugins/fractal2d/index.html.

Another possibility is to examine the suitability of pattern recognition techniques such as the ones presented in [84] or the Fourier analysis [85].

The work presented here provides an insight on the similarity between manual and computational curation of biofilms through clustering. This is an important preliminary step in the development of a workflow for an automated clinical decision. However, a larger, and more heterogeneous image dataset, would add more weight to the conclusions, a clear relationship between colony morphology and the degree of pathogenicity are also required before the development of a decision making software can be implemented in a clinical setting.

References

1. Donlan RM. Biofilms: microbial life on surfaces. Emerg. Infect. Dis. 2002; 8:881–90

2. Pamp SJ, Sternberg C, Tolker-Nielsen T. Insight into the microbial multicellular lifestyle via flow-cell technology and confocal microscopy. Cytom. Part A J. Int. Soc. Anal. Cytol. 2009; 75:90–103

3. Donlan R, Costerton J. Biofilms: survival mechanisms of clinically relevant microorganisms. Clin. Microbiol. Rev. 2002; 15:167–193

4. Renslow R, Lewandowski Z, Beyenal H. Biofilm image reconstruction for assessing. 2012; 108:1383–1394

5. Sommerfeld Ross S, Reinhardt JM, Fiegel J. Enhanced analysis of bacteria susceptibility in connected biofilms. J. Microbiol. Methods 2012; 90:9–14

6. I KM, Jin X, Arne H, et al. An analytical tool-box for comprehensive biochemical, structural and transcriptome evaluation of oral biofilms mediated by mutans streptococci. J. Vis. Exp. 2011; (47). pii:1–9

7. Xavier JB. Ontogenesis of microbial aggregates. 2002;

8. Flemming H-C, Wingender J. The biofilm matrix. Nat. Rev. Microbiol. 2010; 8:623–33

9. Azevedo NF, Lopes SP, Keevil CW, et al. Time to “go large” on biofilm research: advantages of an omics approach. Biotechnol. Lett. 2009; 31:477–85

10. Neu TR, Manz B, Volke F, et al. Advanced imaging techniques for assessment of structure, composition and function in biofilm systems. Fems Micriobiology Ecol. 2010; 72:1–21

11. Simões M, Simões LC, Vieira MJ. A review of current and emergent biofilm control strategies. LWT - Food Sci. Technol. 2010; 43:573–583

12. Cunningham AB, Lennox JE, Ross RJ. Biofilms: The Hypertextbook. 2001-2008

13. Heydorn A, Nielsen A. Quantification of biofilm structures by the novel computer program COMSTAT. Microbiology 2000; 2395–2407

14. Schillinger C, Petrich A, Lux R, et al. Co-localized or randomly distributed? Pair cross correlation of in vivo grown subgingival biofilm bacteria quantified by digital image analysis. PLoS One 2012; 7:e37583

15. López D, Vlamakis H, Kolter R. Biofilms. Cold Spring Harb. Perspect. Biol. 2010; 2:

16. Celiker H, Gore J. Cellular cooperation: insights from microbes. Trends Cell Biol. 2013; 23:9–15

17. Holcombe LJ, O'Gara F, Morrissey JP. Implications of interspecies signaling for virulence of bacterial and fungal pathogens. Future Microbiol. 2011; 6:799–817

18. Albuquerque P, Casadevall A. Quorum sensing in fungi–a review. Med. Mycol. 2012; 50:337–45

19. Bandara HMHN, Lam OLT, Jin LJ, et al. Microbial chemical signaling: a current perspective. Crit. Rev. Microbiol. 2012; 38:217–49

20. Skandamis PN, Nychas G-JE. Quorum sensing in the context of food microbiology. Appl. Environ. Microbiol. 2012; 78:5473–82

21. Wilmes P, Simmons SL, Denef VJ, et al. The dynamic genetic repertoire of microbial communities. FEMS Microbiol. Rev. 2009; 33:109–32

22. VerBerkmoes NC, Denef VJ, Hettich RL, et al. Systems biology: Functional analysis of natural microbial consortia using community proteomics. Nat. Rev. Microbiol. 2009; 7:196–205

23. McDonald D, Vázquez-Baeza Y, Walters W a., et al. From molecules to dynamic biological communities. Biol. Philos. 2013; 241–259

24. Subramoni S, Nguyen DT, Sokol P a. Burkholderia cenocepacia ShvR-regulated genes that influence colony morphology, biofilm formation, and virulence. Infect. Immun. 2011; 79:2984–2997

25. Laanto E, Bamford JKH, Laakso J, et al. Phage-driven loss of virulence in a fish pathogenic bacterium. PLoS One 2012; 7:

26. Allegrucci M, Sauer K. Characterization of colony morphology variants isolated from Streptococcus pneumoniae biofilms. J. Bacteriol. 2007; 189:2030–8

27. Favel A, Michel-Nguyen A, Peyron F, et al. Colony morphology switching of Candida lusitaniae and acquisition of multidrug resistance during treatment of a renal infection in a newborn: case report and review of the literature. Diagn. Microbiol. Infect. Dis. 2003; 47:331–339

28. Zarraonaindia I, Smith DP, Gilbert J a. Beyond the genome: community-level analysis of the microbial world. Biol. Philos. 2012; 28:261–282

29. Spiegelman D, Whissell G, Greer CW. REVIEW / SYNTHÈSE A survey of the methods for the characterization of microbial consortia and communities. 2005; 386:355–386

30. Branda SS, Vik S, Friedman L, et al. Biofilms: the matrix revisited. Trends Microbiol. 2005; 13:20–6

31. Sternberg C, Tolker-Nielsen T. Growing and analyzing biofilms in flow cells. Curr. Protoc. Microbiol. 2006; Chapter 1:

32. Christensen BB, Sternberg C, Andersen JB, et al. Molecular tools for study of biofilm physiology. Methods Enzymol. 1999; 310:20–42

33. Chávez de Paz LE. Image analysis software based on color segmentation for characterization of viability and physiological activity of biofilms. Appl. Environ. Microbiol. 2009; 75:1734–9

34. De Carvalho C, da Fonseca MM. Assessment of three-dimensional biofilm structure using an optical microscope. Biotechniques 2007; 42:616–620

35. Xavier JB, White DC, Almeida JS. Automated biofilm morphology quantification from confocal laser scanning microscopy imaging. Water Sci. Technol. 2003; 47:31–37

36. Mueller LN, De Brouwer JF, Almeida JS, et al. Analysis of a marine phototrophic biofilm by confocal laser scanning microscopy using the new image quantification software PHLIP. BMC Ecol. 2006; 6:1

37. Yang X, Beyenal H, Harkin G, et al. Quantifying biofilm structure using image analysis. J. Microbiol. Methods 2000; 39:109–119

38. Yang X, Beyenal H, Harkin G, et al. Evaluation of biofilm image thresholding methods. Water Res. 2001; 35:1149–58

39. O'Toole GA, Gibbs KA, Hager PW, et al. The global carbon metabolism regulator Crc is a component of a signal transduction pathway required for biofilm development by Pseudomonas aeruginosa. J. Bacteriol. 2000; 182:425–31

40. Pratt LA, Kolter R. Genetic analysis of Escherichia coli biofilm formation: roles of flagella, motility, chemotaxis and type I pili. Mol. Microbiol. 1998; 30:285–93

41. Prigent-Combaret C, Vidal O, Dorel C, et al. Abiotic surface sensing and biofilm-dependent regulation of gene expression in Escherichia coli. J. Bacteriol. 1999; 181:5993–6002

42. Dalton HM, Poulsen LK, Halasz P, et al. Substratum-induced morphological changes in a marine bacterium and their relevance to biofilm structure. J. Bacteriol. 1994; 176:6900–6

43. Davies DG, Geesey GG. Regulation of the alginate biosynthesis gene algC in Pseudomonas aeruginosa during biofilm development in continuous culture. Appl. Environ. Microbiol. 1995; 61:860–7

44. Brooun A, Liu S, Lewis K. A dose-response study of antibiotic resistance in Pseudomonas aeruginosa biofilms. Antimicrob. Agents Chemother. 2000; 44:640–6

45. Costerton JW, Stewart PS, Greenberg EP. Bacterial biofilms: a common cause of persistent infections. Science 1999; 284:1318–22

46. Nielsen AT, Tolker-Nielsen T, Barken KB, et al. Role of commensal relationships on the spatial structure of a surface-attached microbial consortium. Environ. Microbiol. 2000; 2:59–68

47. Møller S, Pedersen AR, Poulsen LK, et al. Activity and three-dimensional distribution of toluene-degrading Pseudomonas putida in a multispecies biofilm assessed by quantitative in situ hybridization and scanning confocal laser microscopy. Appl. Environ. Microbiol. 1996; 62:4632–40

48. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans. Syst. Man Cybern. 1973; SMC-3, N◦6:610–621

49. Neu T. In situ cell and glycoconjugate distribution in river snow studied by confocal laser scanning microscopy. Aquat. Microb. Ecol. 2000; 21:85–95

50. Suraruksa B, Nopharatana A, Chaiprasert P, et al. Microbial activity of biofilm during start-up period of anaerobic hybrid reactor at low and high upflow feeding velocity. Water Sci. Technol. 2003; 48:79–87

51. CA S, WS R. NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 2012; 9:671–675

52. Hartig S. Basic image analysis and manipulation in ImageJ. Curr Protoc Mol Biol 2013;

53. Lamprecht MR, Sabatini DM, Carpenter AE. CellProfiler: free, versatile software for automated biological image analysis. Biotechniques 2007; 42:71–75

54. Daims H, Lücker S, Wagner M. daime, a novel image analysis program for microbial ecology and biofilm research. Environ. Microbiol. 2006; 8:200–13

55. Beyenal H, Donovan C, Lewandowski Z, et al. Three-dimensional biofilm structure quantification. J. Microbiol. Methods 2004; 59:395–413

56. Heydorn A, Ersbøll BK, Hentzer M, et al. Experimental reproducibility in flow-chamber biofilms. Microbiology 2000; 146:2409–15

57. Ferreira T, Rasband W. ImageJ User Guide. 2012;

58. Hope CK, Wilson M. Measuring the thickness of an outer layer of viable bacteria in an oral biofilm by viability mapping. J. Microbiol. Methods 2003; 54:403–10

59. Barraud N, Hassett DJ, Hwang S-H, et al. Involvement of nitric oxide in biofilm dispersal of Pseudomonas aeruginosa. J. Bacteriol. 2006; 188:7344–53

60. Harrison JJ, Ceri H, Yerly J, et al. The use of microscopy and three-dimensional visualization to evaluate the structure of microbial biofilms cultivated in the Calgary Biofilm Device. Biol. Proced. Online 2006; 8:194–215

61. Geissmann Q. OpenCFU, a new free and open-source software to count cell colonies and other circular objects. PLoS One 2013; 8:

62. Selinummi J, Seppälä J, Yli-Harja O, et al. Software for quantification of labeled bacteria from digital microscope images by automated image analysis. Biotechniques 2005; 39:859–863

63. Claytor K. Development and Implementation of an Efficient Automated Cell Colony and Plaque Counter. Univ. Illinois Intern. Phys. Publ. 2008; 7–10

64. Lopes B, Coimbra M. CellNote: A Framework for Assisted Annotation of Cellular Imaging. 2013;

65. Chantratita N, Wuthiekanun V, Boonbumrung K, et al. Biological relevance of colony morphology and phenotypic switching by Burkholderia pseudomallei. J. Bacteriol. 2007; 189:807–17

66. Sousa AM, Lourenço A, Pereira MO. MorphoCol: a powerful tool for the clinical profiling of pathogenic bacteria. Adv. Intell. Soft Comput. 2012; 154:181–188

67. Tajima R, Kato Y. Comparison of threshold algorithms for automatic image processing of rice roots using freeware ImageJ. F. Crop. Res. 2011; 121:460–463

68. Sutherland J. Terminology. 2002;

69. Chinga G, Dougherty B. Roughness Calculation. 2002;

70. Ordonez C, Omiecinski E. Image mining: A new approach for data mining. 1998; 1–22

71. North MA. Data Mining for the Masses. 2012;

72. M.L C, G.T R. Data mining, Classification and Clustering with Morphological features of Microbes. Int. J. Comput. Appl. 2012; 52:1–5

73. Welter P, Riesmeier J, Fischer B, et al. Bridging the integration gap between imaging and information systems: a uniform data concept for content-based image retrieval in computer-aided diagnosis. J. Am. Med. Inform. Assoc. 2011; 18:506–10

74. Shamir L, Delaney JD, Orlov N, et al. Pattern recognition software and techniques for biological image analysis. PLoS Comput. Biol. 2010; 6:

75. Frank E, Hall M, Trigg L, et al. Data mining in bioinformatics using Weka. Bioinformatics 2004; 20:2479–81

76. Zaki MJ, Meira Jr. W. Data Mining and Analysis: Fundamental Concepts and Algorithms. 2013;

77. Van Rijsbergen CJ. Information Retrieval. 1979;

78. Rokach L, Maimon O. Clustering Methods. Data Min. Knowl. Discov. Handb. 2005; 321–352

79. Larsen B, Aone C. Fast and effective text mining using linear-time document clustering. Proc. fifth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '99 1999; 16–22

80. Azevedo NF, Lourenço A, Pereira O. MIABiE - Minimum Information about a Biofilm Experiment. 2011;

81. Hartigan JA. Clustering algorithms. 1975;

82. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. J. R. Stat. Soc. 1979; 28:100–108

83. MacQueen JB. Some Methods for classification and Analysis of Multivariate Observations. Proc. 5th Berkeley Symp. Math. Stat. Probab. 1967; 281–297

84. Ahmed W, Bayraktar B, Bhunia A. Classification of Bacterial Contamination Using Image Processing and Distributed Computing. IEEE J. Biomed. Heal. INFORMATICS 2012;

85. Crawford EC, Mortensen JK. An ImageJ plugin for the rapid morphological characterization of separated particles and an initial application to placer gold analysis. Comput. Geosci. 2009; 35:347–359