Universität Bielefeld

Image Analysis Methods for Location Proteomics

Dissertation zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)

der Technischen Fakultät der Universität Bielefeld

vorgelegt von
Marko Tscherepanow

vorgelegt am 3. Dezember 2007
verteidigt am 5. Mai 2008

Gutachter:

Prof. Dr.-Ing Franz Kummert
Juniorprof. Dr.-Ing. Tim Wilhelm Nattkemper

Prüfungsausschuss:

Prof. Dr. Robert Giegerich
Prof. Dr.-Ing Franz Kummert
Juniorprof. Dr.-Ing. Tim Wilhelm Nattkemper
Dr.-Ing. Thoralf Töpel

Ausdruck der vom Prüfungsausschuss genehmigten Fassung

gedruckt auf alterungsbeständigem Papier nach ISO 9706

Danksagung

Ich möchte hier die Gelegenheit ergreifen, all denjenigen zu danken, die zum Gelingen meiner Dissertation beigetragen haben. Ohne ihre Unterstützung wäre die Fertigstellung dieser Arbeit nicht möglich gewesen.

Mein ganz besonderer Dank gilt dabei meinem Betreuer Franz Kummert, der mir während zahlreicher Diskussionen immer mit seinem Rat zur Seite stand. Mit seinem äußerst umfangreichen Wissen aus allen Bereichen der Mustererkennung eröffnete er mir neue Sichtweisen auf die im Rahmen meiner Arbeit aufgetretenen Probleme. Weiterhin möchte ich ihm und Gerhard Sagerer dafür danken, dass sie mir die Chance zur Bearbeitung dieses spannenden Themas gegeben haben, obwohl mein Vorwissen auf dem Gebiet der Bioinformatik und Proteomik anfänglich sehr eingeschränkt war.

Mein Dank gilt auch Nickels Jensen, Dieter Kapp, Karsten Niehaus und Robert-André Roszik von der Fakultät für Biologie. Durch sie erhielt ich essentielle Einblicke in die zugrunde liegenden biologischen Vorgänge und Mikroskopietechniken, die meine eigene Arbeit wesentlich bestimmt haben. Die Zusammenarbeit mit ihnen hat mir sehr viel Spaß gemacht.

Zusätzlich danke ich allen Mitgliedern der Arbeitsgruppe Angewandte Informatik für ihr Verständnis und ihre Mithilfe; die Berechnungen, die ich im Rahmen dieser Dissertation durchführte, erforderten die Nutzung der meisten ihrer Computer, was deren Leistung teilweise erheblich einschränkte. Auch die äußerst angenehme und kreative Atmosphäre in der Arbeitsgruppe war und ist für mich von sehr großer Bedeutung. Im Besonderen möchte ich hier Frank Zöllner danken, der zu Beginn meiner Promotionszeit mit mir zusammenarbeitete und mir das Verständnis grundlegender Verfahren der Bioinformatik erheblich erleichterte. Außerdem möchte ich mich bei den studentischen Hilfskräften Matthias Hillebrand, Dominik Mertens, Andrea Papst und Tim Nelißen bedanken, die mir zur Seite standen.

Meiner Verlobten Antje Heidbrede, meiner Familie und meinen Freunden möchte ich ebenfalls ganz besonders für ihre Unterstützung, ihr Verständnis und ihre Geduld danken. Ohne Euch hätte ich das nie geschafft!

Abschließend möchte ich mich bei den Korrekturlesern Nils Hofemann, Nickels Jensen und Joachim Schmidt bedanken. Sie trugen wesentlich zur Verständlichkeit meiner Dissertation bei.

Marko Tscherepanow
August 2008

Contents

1 Introduction

The hereditary information or rather the *genome* of several organisms has been read during the last few years. Examples are the nematode *Caenorhabditis elegans* [219], the fruit fly *Drosophila melanogaster* [1], the domestic dog [128] and even humans [100]. But although we have read the genetic message, we do not know its meaning. The hereditary information is transformed by various mechanisms yielding a great variety of components required for building single cells and complex organisms. Here, macromolecules called proteins play a major role.

> *"Proteins are the major components of living organisms and constitute more than 25% by weight of a typical cell. Even more impressive is the variety of functions that they can perform: catalysis, immune recognition, cell adhesion, signal transduction, sensor capabilities, transport, movement, and cellular organization."*
>
> Anna Tramontano (2005) [225, Introduction]

The term 'protein' itself was coined by the Swedish chemist Jöns Jacob Berzelius in the first half of the 19th century [72]. Berzelius derived it from the Greek word *proteios* meaning 'primitive', which highlights the proteins' relevance as elementary building blocks, since they are the basis for constructing living organisms. In order to illustrate this relationship, Chapter 2 introduces the layout of cells as well as their main constituents. Then, the characteristics of proteins, their synthesis and their transport are addressed. Each cell compartment comprises a special set of proteins.

Unfortunately, the functions and interactions of the vast majority of proteins are unknown. The knowledge of these functions could provide crucial information for the simulation of cell behaviour which might facilitate the investigation of diseases as well as the development of innovative drugs and vaccines [29][178]. But the functions of proteins cannot be determined easily, since they can hardly be deduced solely based on the genome [137]. Hence alternative methods enabling the large-scale analysis of proteins have become necessary. They are summarised by the term *proteomics*. The most important proteomic techniques are outlined in Chapter 3.

Chapter 3 starts with fundamental proteomic techniques that enable the separation and identification of single proteins. Furthermore, procedures for the determination of their amino acid sequences and molecular structures are sketched. Such approaches have yielded a vast amount of data. As a result, computational techniques have been developed which aim at extracting new or higher-level knowledge – in particular of a protein's function or interactions – from the available information [71][279]. Their main advantage consists in their significantly increased speed in comparison to laboratory work. But their accuracy is often limited. Therefore, experiments conducted in a biological laboratory are inevitable today.

There are several techniques dealing with this problem. One kind of very promising approach is subsumed under the term *location proteomics* [36]. This branch of proteomics analyses the location of proteins in order to derive information about their function. As specific cell compartments fulfil specific tasks, a protein under consideration is likely to be involved in performing the function of the cell compartments it occurs in.

Since the locations of proteins depend on the cell state and the environmental conditions, they are highly dynamic. Therefore, it is beneficial to examine them in living cells. This can be realised by means of fluorescence microscopy [214]; the proteins in question are tagged by a fluorescent dye and subsequently rendered visible using a fluorescence microscope. Even co-localisations and temporary interactions of multiple proteins become observable. Therefore, location proteomics could reveal crucial details of cell processes. As it is amenable to automatic large-scale analysis, it is of the utmost importance for future proteomic research.

The basic tool of location proteomics consists in microscopy. Without a working knowledge on the applicable microscopy techniques, the characteristics, problems and limitations of this technique cannot be evaluated; in particular, since a fully automated application is necessitated due to the extremely high number of existing proteins. Therefore, fundamental microscopy techniques are reviewed in Chapter 4.

But even if tagged proteins can be observed in live cells, their analysis is difficult. A fluorescence image typically contains multiple or even numerous cells. These cells might be in various states resulting in different locations of the proteins. Furthermore, the cells themselves are not necessarily visible. So a fluorescence micrograph contains bright spots corresponding to accumulations of tagged proteins. These spots vary in size and shape. But they cannot be associated with specific cells and locations therein. A trained biologist might be able to estimate the position of the surrounding cells. However, in a fully automatic context additional information is required. This knowledge can be acquired by employing alternative microscopy techniques in parallel, for instance, bright-field microscopy. Here, the term ‘in parallel’ means that the principal optical path from a specimen to the camera remains unchanged and both images are taken in quick succession. Provided a cell has been recognised in the bright-field image, the corresponding image region of the fluorescence micrograph can be examined. So an association of the observed protein distribution with a specific type of cell compartment becomes feasible. Figure 1.1 illustrates this process.

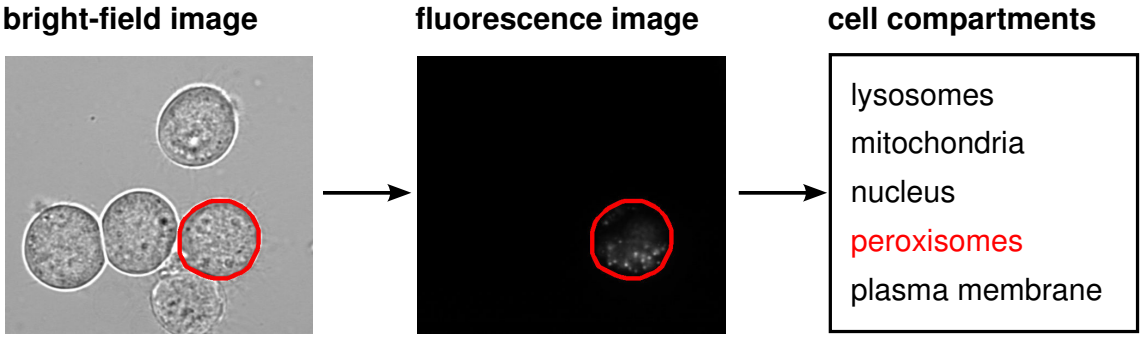


Figure 1.1: *An approach to the automatic localisation of tagged proteins.* A cell (red contour) has been recognised in a bright-field microscope image. The corresponding image region is transferred to the corresponding fluorescence micrograph, where it is analysed. The analysis yields the name of the cell compartment that the tagged proteins are localised in. The remaining cells are examined in an identical way.

In order to enable the analysis of an individual cell’s protein distribution, the region belonging to the cell must be determined; i.e., if a cell has been detected at a specific position, its exact boundaries have to be identified. The process yielding the image regions corresponding to single cells is termed *cell recognition* within the scope of this thesis. This cell recognition provides considerably more information than a cell detection approach, which is sufficient for alternative tasks such as measuring the cell density.

The goal of my thesis consists in providing methods that facilitate the automatic localisation of proteins in living cells. This includes the development of novel cell recognition methods as well as the investigation of mechanisms allowing for an analysis of regions in fluorescence micrographs regarding the distribution of tagged proteins. Furthermore, the combination of both techniques leading to a fully automated system constitutes an innovative contribution to the current proteomic research, in particular, as the focus lies on incremental approaches, which enable the integration of new information during the application.

Chapter 5 addresses the task of cell recognition; a procedure is introduced, which is able to recognise one type of insect cells in bright-field microscope images. This technique can be applied in conjunction with protein localisation approaches without any problems. In order to assess its usability concerning different cell types, it was modified to recognise cells originating from another insect – the fruit fly. Even though these cells impose further problems, the recognition results show its applicability within the context of protein localisation. But in contrast to the original approach, an additional microscopy technique is employed here.

In Chapter 6, methods allowing for the examination of protein location patterns within the recognised cells are discussed. These methods are designed in such a way that the combined application with the proposed cell recognition approach is alleviated. So, a fully automated system is created. Besides considering the task of cell recognition, the focus is on protein localisation techniques, which can be extended to enable retraining during their application. So a biological expert is able to incorporate new information if required.

Eventually, the most important results of my work are summarised in Chapter 7. In addition, the introduced techniques are assessed with respect to the current proteomic research. In particular, the benefit for potential users is discussed. Furthermore, possible extensions are addressed, which could be of interest in the future.

2 Biological Background

In order to understand the importance of proteins for our existence, the structure of living organisms must be considered. The fundamental units of all life forms on our planet are cells. Due to the variety of life, there are essential differences between the existing cell types. Some cells are capable of performing all functions required for living as single organisms, whereas others fulfil very specific tasks as part of a more complex organism. Therefore, some substructures occur only in higher organisms, while others can be found in virtually every cell. These differences affect the biochemical constituents of cells as well. In order to enable a better understanding of these relations, Section 2.1 introduces the main substructures of cells and their most important chemical constituents.

Proteins, a class of very complex macromolecules, are the largest group of biochemical constituents occurring in living cells. Their capability to fulfil specific tasks depends not only on their basic components but also on their three-dimensional structure. Hence, it is crucial to consider the proteins' structure at different levels of detail. The principal structural levels of proteins, which must be taken into account, are outlined in Section 2.2.

The biosynthesis of proteins is reviewed in Section 2.3. This is an intricate process involving several kinds of macromolecules. Hereditary information is translated into functional proteins which allow for the construction of more complex biological structures such as cells.

Eventually, the proteins must be transported to the respective cell compartments to fulfil their task. This process is briefly discussed in Section 2.4.

2.1 Constituents of Living Cells

This Section 2.1.1 introduces the major classes of cells as well as their principal structure. Besides their organelles, living cells comprise various kinds of macromolecules; each of them playing a crucial role for the survival of the cell. The most important macromolecules are reviewed in Section 2.1.2.

2.1.1 Structure of Living Cells

Cells are a biological structure which is separated from its environment by the *plasma membrane* or *cell membrane* [133, Chapter 2]. The interior of a cell is called *cytoplasm*. The fluid is also referred to as *cytosol*. In general, cells must be divided into two major groups: *prokaryotic* and *eukaryotic* cells. Both types differ in size and complexity. In contrast to eukaryotic cells, prokaryotic cells do not possess a membrane-enclosed nucleus. Furthermore, prokaryotic cells are often smaller and they usually lack membrane-enclosed organelles.

The structural similarities of cells were exploited to analyse evolutionary relationships (cf. Section 3.3.1) [133, Chapter 2]. These relationships can be summarised by the phylogenetic tree consisting of three major domains. Prokaryotes comprise the evolutionary domains of the *Bacteria* and the *Archaea*. In contrast, eukaryotic cells constitute an own domain: the *Eukarya*. They encompass all higher organisms such as animals, fungi and plants because of which they are the

preferred cell type for the analysis of protein locations. Figure 2.1 depicts the structure of such a eukaryotic cell.

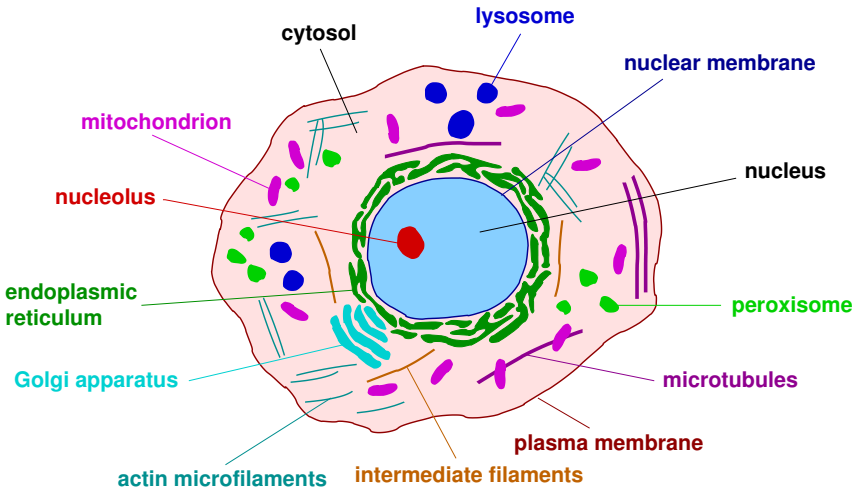


Figure 2.1: *Structure of a eukaryotic cell.* The cell comprises several organelles which are enclosed by an own membrane. Depending on the cell type, differences in the structure might occur. So, for example, plant cells have additional cell walls located outside the plasma membrane in order to increase the structural strength.

The cytoplasm comprises numerous cell compartments performing different functions. The *mitochondria* are responsible for supplying the cell with energy by means of oxidative phosphorylation; adenosine triphosphate (ATP), which constitutes the main energy source of cells, is produced from adenosine diphosphate (ADP) and inorganic phosphate. This process is often referred to as *cellular respiration* [60].

The *Lysosomes* perform *cellular digestion*, i.e. the degradation of macromolecules such as fats, proteins and polysaccharides. The resulting products are passed to the cytoplasm and serve as nutrients for the cell [133, Chapter 14].

Peroxisomes are containers for enzymes required for oxidative reactions [167, Chapter 1]. They produce toxic hydrogen peroxide, which is degraded to oxygen and water.

The structure of eukaryotic cells is reinforced by the *cytoskeleton*. The necessity of reinforcement results from the large size of eukaryotic cells as well as their motility [133, Chapter 14]. It consists of filamentous structures called *actin microfilaments*, *intermediate filaments*, and *microtubules*.

Apart from the cytoskeleton, several organisms possess *cell walls* in order to reinforce their cell structure [133, Chapter 2]. Cell walls occur in prokaryotes as well as in eukaryotes and are situated outside the cell membrane. They are permeable so that the exchange of substances with the environment is not blocked.

The *genome*, i.e. the hereditary information of a eukaryotic cell, is mainly contained in the *nucleus* [133, Chapter 14]. The nucleus is separated from the cytoplasm by the *nuclear membrane*, which actually comprises a pair of membranes. The inner membrane realises interactions with the interior of the nucleus, whereas the outer membrane is specialised on interactions with the cytoplasm. The latter is directly connected to the *endoplasmic reticulum* (ER). The ER constitutes a system of tubules and stacks which is involved in the synthesis of proteins as well as their transport [60]. The ER is connected to the nuclear membrane. In addition, it is able to transport its products to the Golgi apparatus where they are chemically modified [133, Chapter 14].

The nucleus encloses structures called *nucleoli*. The nucleoli are involved in the production of *ribosomes*, which are the sites of protein synthesis (see Chapter 2.3) [60].

Chloroplasts constitute another very important cell compartment. They appear in eukaryotes like algae and plants that are able to perform photosynthesis [133, Chapter 14]; that is, the production of carbohydrates from carbon dioxide and water using sun-light [60].

Mitochondria and chloroplasts contain an own genome, though the majority of their functions are encoded in the nucleus. Furthermore, they are able to synthesise own proteins. Their structure resembles prokaryotes. Therefore, it is assumed that both of them originate from prokaryotes which were included in eukaryotic cells [133, Chapter 14].

2.1.2 Macromolecules

Approximately 96% of the dry weight of cells comprises *macromolecules* [133, Chapter 3]. Macromolecules are large molecules composed of smaller units which are called *monomers*. Since they encompass several monomeric units, they are referred to as *polymers*, as well. In principle, three important classes of macromolecules must be distinguished: proteins, nucleic acids and polysaccharides.

Polysaccharides are the smallest group of macromolecules which can be found in a cell. These polymers constitute large chains of carbohydrates or rather sugars. In living cells, they serve as carbon and energy reserves, or as reinforcing material in order to form the cell wall. In addition to sugars, polysaccharides can contain other molecules such as proteins and *lipids*. The lipids of Bacteria and Eukarya consist of fatty acids. As fatty acids comprise water-repelling (*hydrophobic*) and water-soluble (*hydrophilic*) regions, they are beneficial for constructing permeable barriers such as the cell membrane.

Nucleic acids emerging in living cells can essentially be divided into two types: *deoxyribonucleic acid* (DNA) and *ribonucleic acid* (RNA). DNA essentially encodes the hereditary information of a cell, whereas RNA fulfils various tasks. Besides serving as intermediary information source for the synthesis of proteins (see Section 2.3), it may have structural or catalytic functions [133, Chapter 3].

Both types of nucleic acids are constructed from monomers called *nucleotides*. These nucleotides comprise three components: a sugar (ribose or deoxyribose), a nitrogenous base and a phosphate molecule. Here, five different nitrogenous bases are utilised: *adenine*, *guanine*, *thymine*, *cytosine* and *uracil*. While thymine exclusively occurs in DNA, uracil is only present in RNA. Apart from their usage for DNA and RNA, nucleotides might have a function of their own. So, for example, the nucleotide adenosine triphosphate is employed as an energy source of a cell (cf. Section 2.1.1).

The two-stranded structure of DNA was discovered by James D. Watson and Francis H. C. Crick at the University of Cambridge in 1953 [256]. Both strands may comprise several million nucleotides. The most stable bond between both stands occurs if they are complementary, that is, each guanine is opposite of a cytosine and every adenine opposite of a thymine. Therefore, the ratio of adenine and thymine as well as the ratio of guanine and cytosine equal one, as it was first stated by the Austrian-born biochemist Erwin Chargaff in 1951 [31]. The strands of DNA are positioned antiparallel and form a right-handed double helix, which is referred to as its *secondary structure* [167, Chapter 2] (see Figure 2.2).

In contrast to DNA, RNA is usually single-stranded. But it can fold back on itself in order

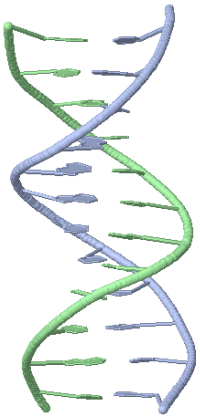


Figure 2.2: *Secondary structure of DNA*. DNA comprises two strands arranged in the form of a right-handed double helix. The bases are schematically illustrated. Here, bases being situated opposite each other are complementary.

to form bonds between complementary bases. The corresponding pattern is called the secondary structure of the RNA. Besides serving as intermediary information source, some RNAs called *ribozymes*, have a catalytic function. It is assumed that these ribozymes are remains from ancient precursors of living organisms [133, Chapter 7]. Nevertheless, they have important functions, for example in protein synthesis (see Section 2.3). Such an RNA's secondary structure is crucial for its activity.

Proteins constitute the largest group of macromolecules occurring in living cells. They comprise, for instance, about 55% of the dry weight of a prokaryotic cell [133, Chapter 3]. In a single mammalian liver cell there are approximately eight billion protein molecules [141, Chapter 4]. With respect to their function, two classes can be distinguished: *enzymes* and structural proteins. Enzymes function as *catalysts* for numerous chemical reactions, i.e., they facilitate the reaction by decreasing its activation energy [133, Chapter 5]. But they are neither consumed nor otherwise affected by the reaction. In contrast, structural proteins are of crucial importance for constructing cell walls and cell organelles.

A protein consists of one or more polymers, the *polypeptides* [167, Chapter 2]. The monomeric units of such a polypeptide are referred to as *amino acids* (see Section 2.2.1). Amino acids, which are linked in polypeptides, are called *residues*. The majority of proteins occurring in cells comprise 100 to 1000 residues.

Until 1986 it was assumed that proteins were constructed from 20 standard amino acids encoded by DNA. Then Ian Chambers and his co-researchers discovered a 21st DNA-encoded amino acid: selenocysteine [28]. In 2002, the discovery of a 22nd amino acid called pyrrolysine was reported by Bing Hao and his colleagues [83]. Like selenocysteine and the standard amino acids, pyrrolysine has a DNA representation. Therefore, more DNA-encoded amino acids are likely to be discovered in the future. In addition, amino acids might be altered after protein synthesis (see Section 2.3). So, proteins can contain residues which are not directly encoded by DNA. More than 200 such amino acids are known today [253, Chapter 1].

2.2 Proteins

As explained in Section 2.1.2, proteins are very complex macromolecules which are composed of polymers constructed from amino acids. Therefore, different aspects of their structure must be considered. At first, Section 2.2.1 gives an overview about the structure and properties of amino

acids. Afterwards, Section 2.2.2 explains, how these simple building blocks are combined in order to yield complete proteins. Eventually, the process of protein synthesis is introduced in Section 2.3.

2.2.1 Amino Acids

Amino acids are organic acids which are employed to form proteins and polypeptides. They usually encompass an amino group, a carboxylic acid group and a side chain which are attached to a central carbon atom – the α-carbon (see Figure 2.3) [60]. Depending on the relative position of the amino group and the carboxylic acid group, amino acids are further classified into α-, β-, γ-, and δ-amino acids. Proteins are exclusively composed of α-amino acids. β-, γ-, and δ-amino acids may occur as free acids or components of other organic products.

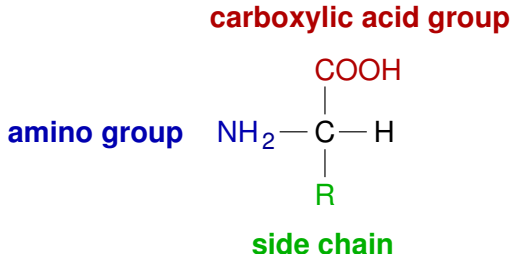


Figure 2.3: *General structure of an α-amino acid.* Both the carboxylic acid group and the amino group are linked with the α-carbon.

The biochemical properties of an amino acid mainly result from its side chain. Due to the side chain, an amino acid may be, for instance, acidic, basic, hydrophobic or hydrophilic. The chemically-active sites as well as the higher-level structures of a polypeptide are almost completely determined by these side chains.

Although having the same molecular formula, α-amino acids might exhibit a different structure. It is possible that two amino acids are mirror images of one another. The corresponding molecules are called *enantiomers* [133, Chapter 3]. The enantiomer usually used for protein synthesis is symbolised by L. D denotes the alternative acid. Nonetheless, D-amino acids are used by some cell wall polymers.

With respect to an organism's ability to synthesise the required amount of a specific amino acid, it is referred to as *essential* and *non-essential*, respectively. Essential amino acids must be provided by nutrition. Otherwise, the production of proteins would be hampered. In contrast, the organism itself can produce non-essential amino acids.

2.2.2 Structure of Proteins

Due to the complexity of proteins, different levels of their structure are usually distinguished [133, Chapter 3][253, Chapter 1]. At first, the sequences of amino acids in each polypeptide are considered. These sequences are referred to as *primary structure*. The primary structure determines the way of folding the final polypeptide. So, it is fundamental for its function. Unfortunately, the prediction of a resulting protein's three-dimensional structure based on its polypeptides' sequences is extremely difficult [167, Chapter 2].

The *secondary structure* of a polypeptide describes the way it is folded. Here, two main types are considered: *α-helices* and *β-sheets* (see Figure 2.4). An α-helix constitutes a linear part of a polypeptide which seems to be wound around a cylinder. In contrast, a β-sheet comprises several flat structures, the *β-strands*, folded parallel or antiparallel to each other. While α-helices exhibit

flexibility to a certain degree, β-sheets are rather rigid. It is not unusual for a polypeptide to contain both kinds of secondary structure. The connections between them are formed by loop regions.

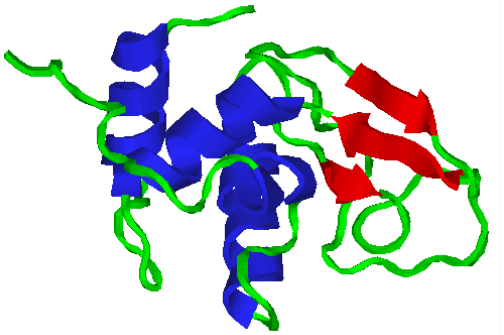


Figure 2.4: *Structure of lysozyme.* Lysozyme is an enzyme weakening the cell walls of Bacteria so as to prevent infections [133, Chapter 4]. Here, a special version found in hen egg white is depicted [185]. The α-helices are shown in blue and the β-sheets in red. Green-coloured loop regions realise their connection.

The relations of α-helices and β-sheets constitute the *tertiary structure* of a polypetide [133, Chapter 3]. As a result of the arrangement of these components, exposed regions and grooves in the molecule's surface are created. These regions frequently take part in fulfilling its function.

The number and type of different polypeptides a protein is composed of are referred to as *quaternary structure*. In principle, proteins may comprise identical or nonidentical subunits. Figure 2.5 demonstrates this by means of the example of hemoglobin. In addition, interacting proteins associate and form *protein complexes*. The ribosomes constitute an example for such complexes. Besides proteins, other molecules, e.g. RNA, can be incorporated as well.

Figure 2.5: *Quaternary structure of human hemoglobin.* Hemoglobin is a protein contained in the red blood cells (erythrocytes) of vertebrates causing the red colour of blood [60]. Its main task consists in the transport of oxygen. Hemoglobin is constructed from four polypetides, which have been coloured differently. All of them have their own primary, secondary and tertiary structure.

The three-dimensional shape of a protein is essential for its biological properties. If the higher-order structures of a protein are destroyed, e.g. by extreme heat, the protein unfolds and loses the ability to fulfil its function, although the primary structure is not necessarily affected. This process is called *denaturation* [133, Chapter 3].

2.3 Protein Biosynthesis

The biosynthesis of proteins, which is frequently termed *expression*, comprises several processes, since DNA is not directly converted into proteins. At first, functional units called *genes* are copied

into RNA which is used as intermediate information source (see Section 2.3.1). The process of transferring information from DNA to RNA is called *transcription* [60][133, Chapter 7]. In eukaryotic cells, it is usually performed in the nucleus where DNA is situated in the form of *chromosomes* which constitute large DNA molecules [167, Chapter 14][60].

The transcribed RNA is not always fully functional. Frequently, it first has to be processed in order to be able to achieve its task (see Section 2.3.2). Especially in higher organisms, *post-transcriptional changes* occur.

After its processing, eukaryotic RNA is transferred to the cytoplasm. There it is employed so as to construct proteins. A small number of proteins is encoded by DNA in mitochondria or chloroplasts. Their synthesis takes place in the respective organelle [167, Chapter 18]. But the majority of the respective compartments' proteins are synthesised in the cytoplasm as well [167, Chapter 19].

In order to construct a protein, the RNA's sequence of nucleotides must be first *translated* into an sequence of amino acids (see Section 2.3.3). But the resulting amino acid chain is not necessarily equal to a final chain of a protein. It frequently has to be processed so as to guarantee particular structural properties or enable its biological activity. Such *post-translational modifications* reach from the chemical alteration of single amino acids up to the cleavage of the nascent polypeptide (cf. Section 2.3.4).

Before the final protein can be assembled, its amino acid sequences need to be folded properly (see Section 2.3.5). Afterwards, the polypeptides can associate.

2.3.1 Transcription

The transcription is mainly performed by an enzyme – the *RNA polymerase*. Prokaryotes utilise a single RNA polymerase while more complex eukaryotes apply three different types; each of them associated with a specific kind of gene. Due to its relatively large size, the RNA polymerase can form contacts with many consecutive DNA bases at the same time. The process of transcription can be divided into three principal steps: initiation, elongation and termination.

At the beginning, in the initiation phase, the RNA polymerase binds at a particular site of the DNA called *promoter*. Such a promoter is characterised by a specific sequence of nucleotides. After binding, the RNA polymerase opens the DNA double helix in a short segment, since only one strand of the DNA is utilised for transcription. The orientation of the promoter's sequence determines the strand which is to be applied.

During the second phase, the elongation, the RNA polymerase moves along this strand, away from the promoter and synthesises RNA simultaneously (see Figure 2.6). The usage of uracil instead of thymine (cf. Section 2.1.2) does not affect the base pairing, since both nitrogenous bases bind equally well with adenine. On average, between 20 and 30 nucleotides are attached to the new RNA strand in one second [167, Chapter 14].

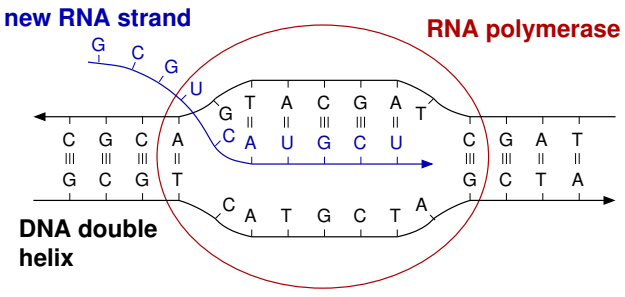


Figure 2.6: *Transcription of DNA into RNA.* The RNA polymerase opens the DNA double helix and creates a copy in RNA. Here, the nitrogenous bases are symbolised by the first letter of their name. Uracil is incorporated into the new RNA strand instead of thymine.

The last phase, the termination, starts when the polymerase reaches a certain sequence called *transcription terminator* [133, Chapter 7]. Here, the RNA strand dissociates from the RNA polymerase and the utilised DNA returns to its original conformation. The new RNA strand is an exact copy of the DNA. But before its application, some modifications might be necessary. In such cases, the transcribed RNA is named *precursor RNA* (pre-RNA).

2.3.2 Post-Transcriptional RNA Processing

Transcribed RNA is frequently processed before it is able to achieve its task, i.e, *mature RNA* is created based on precursor RNA. In eukaryotic cells, the RNA is changed comprehensively; for example, it might be cut, joined or chemically modified [167, Chapter 15]. The RNA occurring in living cells can essentially be divided into three major types: *messenger RNA* (mRNA), *transfer RNA* (tRNA), and *ribosomal RNA* (rRNA); mRNAs contain the blueprint of proteins which are to be synthesised, tRNAs transport specific amino acids, and rRNAs are required by the ribosomes. Each type of RNA is encoded by different genes.

Genes, especially eukaryotic genes encoding for mRNA, contain regions which are not applied for the construction of the final polypeptide [133, Chapter 7]. These noncoding sequences, the *introns*, must be removed after transcription. Then, the gene's coding regions, the *exons*, are joined. Interestingly, the introns may be far larger then the exons. It is assumed that the existence of introns facilitates the evolution, since new genes can be formed by the combination of fractions from existing ones with segments encoded by introns [167, Chapter 15]. The process of removing intros and joining exons is referred to as *splicing*. It is realised by complexes, the *spliceosomes*, which comprise proteins as well as a small RNA.

Intriguingly, splicing can be performed in different ways [167, Chapter 15]. So, multiple RNAs can be formed based on the same gene in a process controlled by enzymes; the introns of one mRNA might be considered as exons of another one. The result are mRNAs with partly overlapping nucleotide sequences. This *alternative splicing* particularly occurs in higher organisms; for example, about a quarter of human mRNAs is spliced alternatively.

Besides splicing, the ends of a pre-mRNA are modified in order to yield the final RNA [167, Chapter 15]. These modifications accomplish numerous functions: They protect the mRNA from degradation, support transportation to the cytoplasm and splicing, and assist the translation process. The complete set of mRNA in a living cell is termed *transcriptome* [236, Chapter 1].

tRNA and rRNA are usually encoded by genes which occur several times. Such genes may have many copies in eukaryotes, whereas prokaryotes possess only a limited number. After transcription, the precursor RNAs of tRNA and rRNA are cut so as to yield the corresponding mature RNAs. In addition, numerous tRNA bases are modified. Therefore, tRNA contains several bases such as inosine which are not genetically encoded [133, Chapter 7]. These modifications are required in order to make the tRNA functional. They do not encode for proteins. rRNA and even DNA can contain modified bases, as well, but in comparison to tRNA they occur less frequently [60].

2.3.3 Translation

When mRNAs have been processed, their nucleotide sequence is *translated* into an amino acid sequence so as to form polypeptides. As four different bases are utilised for encoding 20 standard

amino acids, there does not exist a one-to-one mapping. In fact, three consecutive bases encode for one amino acid. Such a base triplet is called a *codon*. Since, in principle, a triplet of four different bases could encode one of $4^3 = 64$ amino acids, there exist more codons then required. Some of these additional codons indicate the start and the end of translation. A sequence of triplets enclosed by a start codon at the beginning and a stop codon at the end is referred to as *open reading frame*. Besides control codons, several related base triplets encode for the same amino acid. So, there are no unused base triplets.

Unfortunately, the *genetic code*, i.e. the mapping from nucleotides to amino acids, varies between different organisms and organelles [133, Chapter 7]. In particular, the meaning of stop codons changes. So it is possible to encode for two additional amino acids not included in the standard set: selenocysteine [28] and pyrrolysine [83]. Here, the context of the codon is exploited in order to distinguish between both meanings. There are other modifications as well. But all of them resemble the standard code and are therefore assumed to be evolutionarily related to it.

The base triplets encoding for amino acids are read by particular tRNAs. In order to achieve this task, tRNA contains a complementary sequence of the codon in question, the *anticodon* (cf. Figure 2.7). During protein synthesis the anticodon binds at the respective codon. But there is not one tRNA for each codon. Some tRNAs allow for a more flexible base-pairing. Thus, the codon and the anticodon need not match at specific positions. This phenomenon is called *wobble*. Besides the anticodon, tRNA has a substructure that is responsible for binding with the appropriate amino acid – the *acceptor end*. The coupling of an amino acid to its tRNA is controlled by enzymes termed *aminoacyl-tRNA syntheases* [167, Chapter 18].

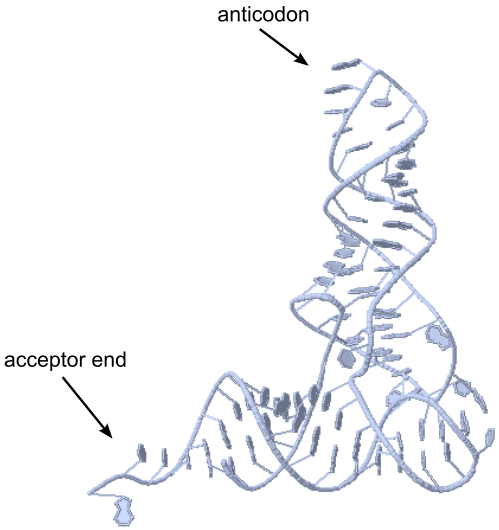


Figure 2.7: *Three-dimensional structure of transfer RNA.* The depicted molecule constitutes the yeast tRNA for the amino acid phenylalanine [107]. It encompasses a total of 76 nucleotides, three of which are forming the anticodon binding at mRNA. Three others constitute the acceptor end. This substructure is responsible for binding at phenylalanine.

The translation takes place in the *ribosomes* – complexes consisting of rRNA and ribosomal proteins [133, Chapter 7]. Each ribosome is composed of a small and a large subunit. They combine the mRNA, which is to be translated, with tRNAs carrying specific amino acids. Similar to the transcription (cf. Section 2.3.1), the translation can be divided into three phases referred to as initiation, elongation and termination. All of these phases are controlled by special proteins.

In the initiation phase, a complex comprising a small ribosome subunit and an initiator tRNA, which corresponds to the start codon, binds with the mRNA to be translated. Subsequently, a large ribosome subunit is attached.

Afterwards, the elongation phase starts. Here, the mRNA is moved in steps of three amino acids. The codons are coupled with the respective tRNAs. In addition, the corresponding amino acids are linked so as to form a polypeptide. It is possible that several ribosomes translate the same strand of mRNA simultaneously. Such a complex is called *polysome*. Polysomes allow for a considerable increment of the translation's efficiency.

Finally, the translation is terminated when a stop codon is reached. In contrast to a start codon, there is no tRNA for a stop codon. So, the formed polypeptide is released instead of coupling a new amino acid to it. Then, the subunits of the ribosome dissociate.

2.3.4 Post-Translational Modification

Every kind of difference of a translated polypeptide sequence with respect to the final functional protein is referred to as *post-translational modification* [60]. These modifications can already occur before the complete polypetide sequence is assembled. The main task of such modifications consists in altering the biological activity or structural properties of a polypeptide [253, Chapter 1]. In principle, three kinds of post-translational modification are distinguished: proteolytic processing, the addition of prosthetic groups and the modification of the amino acids' side chains.

Proteolytic processing describes the specific cleavage of a polypetide chain by cellular enzymes called *proteases*. As a result, the polypeptide might be activated or modified in order to be transported to particular cellular organelles or to be secreted from the cell.

The addition of *prosthetic groups* encompasses the attachment of components which are not constructed of amino acids. Frequently, carbohydrates are added resulting in *glycoproteins*. Glycoproteins are important components of the cell membrane and body fluids. They comprise numerous enzymes, *hormones* enabling a communication between cells of an organism and all *antibodies*, which are an integral part of the *immune system* – the body's defence against infections [60].

Only 22 amino acids are known to be encoded by the genetic code. But more than 100 have been observed in naturally-occurring proteins. The additional amino acids are the product of post-translational modifications of the translated amino acids' side chains. In eukaryotic cells, it is largely performed in the endoplasmic reticulum, the Golgi apparatus, and the cytosol .

2.3.5 Folding

Although, in principle, the three-dimensional structure is determined by their amino acid sequence, many polypetides need assistance in order to fold correctly [133, Chapter 7] [253, Chapter 1]. This task is accomplished by a class of proteins called *molecular chaperones*. As their concentration rises in the case of high temperatures, they were previously called heat-shock proteins. Besides their ability to assist nascent polypeptides, they can help to refold denatured proteins as well. So, proteins unfolded due to heat can be repaired.

One major class of chaperones, the *chaperonins*, occur in virtually every known organism, which emphasises their importance. Chaperonins exhibit a structure resembling a barrel with open ends (see Figure 2.8), which is passed by polypeptides to be fold. After their release, the polypeptides have a proper secondary and tertiary structure.

The final polypeptides are capable of interacting with each other. So, they can form proteins consisting of several polypeptide chains. The complete set of proteins occurring in a cell is referred to as its *proteome*.

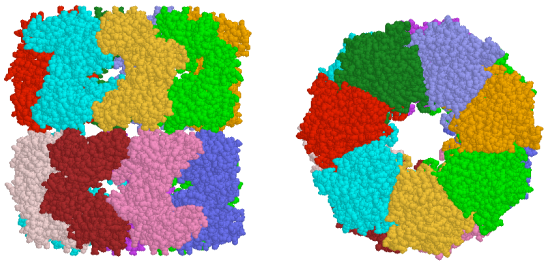


Figure 2.8: *Structure of chaperonins.* Chaperonins such as the bacterial chaperonin GroEL [254], which is depicted here, exhibit a barrel-like shape. This structure is visualised from two different viewpoints. Atoms are depicted as spheres. Moreover, adjoining polypeptide chains were coloured differently.

2.4 Targeting of Proteins

Besides constructing proteins properly, they must be transported to the right location, i.e., they have to be moved from the site of synthesis[1] to the organelles where they are required. The target of a protein is encoded in its amino acid sequence [168, Chapter 19]. The corresponding residues are called *targeting signals*. These residues do not need to be contiguous. Furthermore, they may be removed during transport or be permanent. There are even proteins which contain multiple targeting signals, for example, one signal responsible for transporting the protein to a cell organelle and another one directing it to a subcompartment.

Depending on the target and a protein's function, the process of protein targeting may occur coincidently with protein biosynthesis (cotranslational targeting) or after it (post-translational targeting). Proteins required by the mitochondria, the chloroplasts and the peroxisomes are usually translocated post-translationally [168, Chapter 19]. Other proteins, such as the ones destined for the endoplasmic reticulum, the golgi apparatus, the plasma membrane or secretion may be transported during their synthesis [168, Chapter 20]. Here, the term *secretion* denotes the transport of proteins out of the respective cell.

[1] mainly the cytosol

3 Determination of a Protein's Function

Since proteins are virtually involved in performing every kind of biological function, the examination of the complete set of proteins – the proteome – which is expressed in a specific cell under particular conditions reveals information on how the cell is working [236, Chapter 1]. This knowledge contributes to our understanding of basic biological processes. So, for example the development of innovative remedies and therapies is supported.

In contrast to the genome that does not usually change during the lifetime of a cell, the proteome is modified with respect to its actual condition determined by internal and external factors. So, it is highly dynamic. Furthermore, due to processes such as post-translational modifications (cf. Section 2.3.4), the number of proteins is significantly higher than the number of genes occurring in an organism. The human genome, for example, is assumed to contain about 30,000 genes, whereas the human proteome is likely to encompass more than one million proteins. Therefore, the proteome's analysis is inherently more complex than the analysis of the corresponding genome. As a result, methods allowing for a large number of experiments in a short time, i.e. high-throughput techniques, are required. Such methods enable the investigation of dynamic processes, which were not possible based solely on hereditary information.

The proteome's analysis may be carried out by a variety of methods. The first techniques employed methods available from genome analysis, i.e. *genomics*, and applied them to mRNA [236, Chapter 1]. As the transcriptome was analysed, this kind of investigation has been called *transcriptomics*. Transcriptomics enables the analysis of proteins on a more detailed but nonetheless indirect level; the proteins' structures, post-translational modifications, their locations and their abundances are not taken into account.

Nowadays, the proteome can be examined directly describing the state of a living cell more detailed [155]. Depending on the context, different methods must be employed, which are summarised by the term *proteomics*; for example, features such as a protein's amino acid sequences, structure, abundance, interactions and location might be of interest. Several important techniques enabling these kinds of investigation are outlined in the following sections. In principle, all of them aim at determining the function of proteins. Here, two major problems arise: Firstly, the meaning of the term function is not unambiguous. Rather it might refer to diverse levels and various degrees of detail [225, Problem 3][236, Chapter 5]. Secondly, some proteins, called *moonlight proteins* [225, Chapter 6], fulfil more than one task. So, the same protein might be involved in diverse cellular processes. The actual function is determined by a variety of aspects such as the cell type, its location within the cell and the presence of ligands. Although I do not go into more detail, these two problems must be kept in mind if the function of proteins is considered.

Before discussing individual proteomic techniques, Section 3.1 introduces basic properties of electromagnetic radiation which is a crucial probe utilised by numerous experimental techniques. Afterwards, Section 3.2 addresses fundamental methods for the analysis of proteins. These methods yield information on the amino acid sequence, the structure and the quantity. In addition, they enable individual proteins to be identified and extracted from a mixture, which constitutes a prerequisite for further studies.

By virtue of the large number of experiments that have been conducted in recent years, a large number of amino acid sequences and structures is available. This allows for a transfer of knowledge from well-characterised proteins to others, which share a similar amino acid sequence or whose three-dimensional structures resemble each other. Moreover, conclusions about a protein's function as well as its interactions partners can be drawn based on these data. Basic techniques for such a computational analysis of stored experimental results are addressed in Section 3.3. The main advantage of such techniques consists in the fact that laboratory work is not required. So, a large number of investigations can be performed in a short period of time.

But unfortunately, such analyses do not lead to very accurate results with respect to biological questions. Therefore, additional experimental techniques are necessitated as a means of verification. These procedures are introduced in the subsequent sections. Here, Section 3.4 concentrates on the investigation of protein interactions, which constitute the basis of numerous biological processes [236, Chapter 7].

But information about a proteins' function cannot only be derived from the knowledge of its possible interaction partners. Its localisation might reveal crucial information as well. Additionally, interactions of proteins occurring at the same location within a cell at the same time might be visualised. Therefore, *location proteomics*, i.e. the automatic subcellular localisation of many or all proteins of a cell, has made considerable progress during the last decade [36]. In several aspects it is superior to the approaches mentioned before; for example, it allows for an observation of true protein interactions occurring in their natural environment. The basic methods and application fields of location proteomics are discussed in Section 3.5. In principle, the observation of protein locations is amenable to high-throughput processing. But in order to achieve this goal, new image analysis methods are necessitated. Therefore, the goal of my thesis is to provide methods, which have the potential to facilitate the high-throughput localisation of proteins in the future.

The most important characteristics of the discussed procedures are summarised in Section 3.6. There, the benefits and drawbacks of the considered proteomic approaches are contrasted.

3.1 The Nature of Electromagnetic Radiation

Electromagnetic radiation constitutes a critical means for analysing proteins. It is classified according to its energy. This energy can only take on multiples of discrete units or rather quanta called photons, which exhibit properties of particles as well as of waves. So, a photon can be described by an electric and a magnetic field vibrating perpendicular to one another (see Figure 3.1).

On the other hand, a photon can be considered as a particle possessing an energy E which is derivable from its frequency ν according to Equation 3.1.

$$E = h \cdot \nu \tag{3.1}$$

This relation was discovered by the German physicist Max Planck [164]. Therefore, the constant h is called Planck's constant. Based on the frequency ν, the wavelength λ can be obtained according to Equation 3.2 where c denotes the speed of light.

$$\lambda = \frac{c}{\nu} \tag{3.2}$$

Due to the relations described by Equation 3.1 and Equation 3.2, the wavelength is usually applied as a measure for the energy. The corresponding *electromagnetic spectrum* depicting the

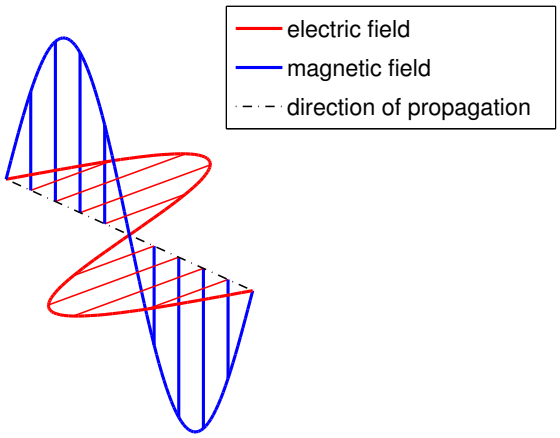


Figure 3.1: *Electromagnetic radiation as a wave.* If considered as a wave, electromagnetic radiation is composed of an electric and a magnetic field perpendicular to each other as well as to the propagation direction.

characteristics of several important types of electromagnetic radiation is shown in Figure 3.2.

According to its ability to remove electrons from atoms or molecules, electromagnetic radiation is divided into *ionising* and *non-ionising radiation*. Besides influencing other molecules, ionising radiation is able to interact with DNA and cause breaks in its structure, which might result in considerable damage of living cells [133, Chapter 20]. It comprises high-energetic cosmic rays, γ-rays, X-rays, and ultraviolet (UV) radiation [80, Chapter 1]. In contrast, visible light as well as infrared and radio frequency radiation are not able to ionise molecules, which makes them suitable for the analysis of live specimens. Furthermore, electromagnetic radiation often comprises a variety of photons. Therefore, its composition becomes important.

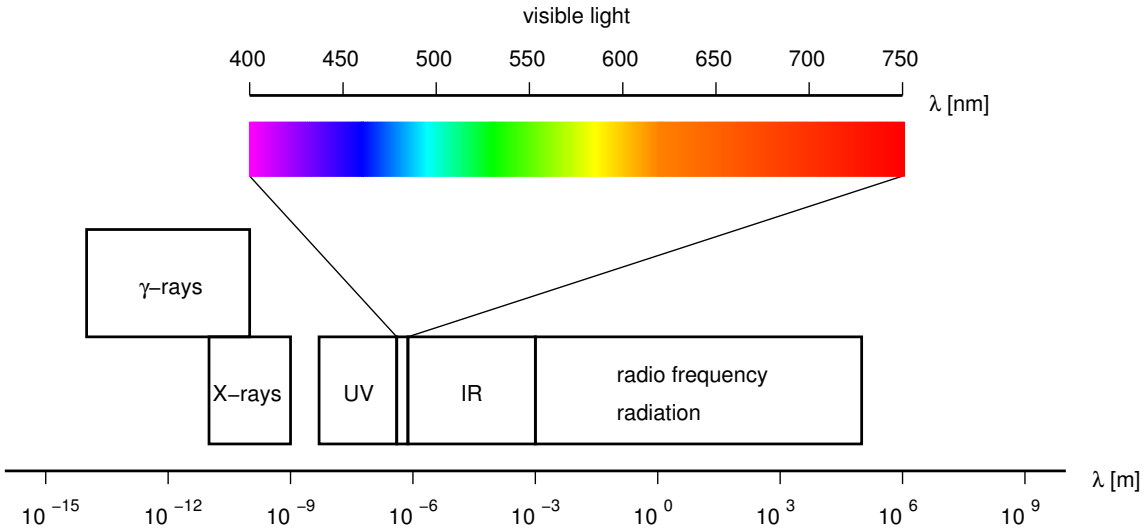


Figure 3.2: *Electromagnetic spectrum.* Electromagnetic radiation occurs in a great variety of different types characterised by their energies which are inverse proportional to their wavelengths λ. Here, visible light covers only a very narrow range; colours are represented by different wavelengths. The direct neighbours of visible light in the spectrum are ultraviolet (UV) radiation, which is more energetic, and infrared (IR) radiation possessing a larger wavelength. (Based on data taken from [80, Chapter 1] and [144, Chapter 2])

Visible light, which is frequently utilised in microscopy, encompasses only frequencies between

400nm and 750nm [144, Chapter 2]. Here, different frequencies correspond to different colours, which are called *spectral colours*.

3.2 Fundamental Proteomic Techniques

In this section, basic proteomic methods enabling the acquisition of fundamental knowledge about proteins are introduced. In order to fulfil this task, methods for the separation of proteins are necessary. Such methods are introduced in Section 3.2.1. Then, individual proteins are identifiable, for example by determining their sequence, which enables further investigations (see Section 3.2.2). Another essential feature of proteins consists in their three-dimensional structure. Therefore, Section 3.2.3 discusses methods capable of solving this structure. Eventually, techniques allowing for a measurement of the protein abundance are addressed in Section 3.2.4. So, changes of the proteome between different cell types or cell states can be quantitated.

3.2.1 Separating Proteins

The separation of protein mixtures is an essential component of proteomic methods enabling investigations concerning individual proteins [236, Chapter 2][45]. One procedure capable of performing this task is termed *two-dimensional gel electrophoresis* (2DGE). The two-dimensional gel electrophoresis is a procedure for separating proteins on a plate called *gel* [236, Chapter 2].

2DGE is usually based on two features: electric charge and mass. At first, an electric field is applied which causes the proteins to migrate in the first dimension depending on their charge. This process is called *electrophoresis*. In fact, the charge is not analysed directly. Instead, the change of a protein's charge with respect to the *pH* of its surrounding is investigated.[1] In order to achieve this goal, a pH gradient is created on the gel. The employed electric field causes the proteins to migrate: positively charged proteins to the cathode and negatively charged proteins to the anode. Since the proteins move along the pH gradient, their charge with respect to their surrounding alters. As a result, they reach a point, where they are not charged at all – the *isoelectric point*. Here, they stop migrating.

Then, the proteins are exposed to sodium dodecyl sulfate (SDS), which constitutes a negatively charged molecule. So they get denatured, i.e. they unfold. Now, the SDS molecules bind to the proteins resulting in a significant increase of the charge. The number of SDS molecules, which bind to a protein, is roughly proportional to its size and mass. Afterwards, a new electric field is applied in the second dimension causing the proteins to migrate again. Here, the gel acts as some kind of sieve decreasing the speed of proteins depending on their size. Eventually, the proteins are stained in order to visualise the protein accumulations on the gel (see Figure 3.3). Automatic evaluation methods for digitised gel images alleviate the investigations and enable high-throughput processing [45][110]. On the other hand, new sources of variability concerning an investigation's outcome may be introduced [259].

Proteins located on two-dimensional gels are collected in the SWISS-2DPAGE database [87]. Due to the high number of proteins contained in living cells, the resolution of the gel is a critical feature. Therefore, several attempts have been made to increase it. One approach called *digital deconvolution* [5], which is relevant for further image-based proteomic techniques, is introduced in Section 4.3.1.

[1] The pH reflects the concentration of hydrogen ions within a solution [133, Chapter 6].

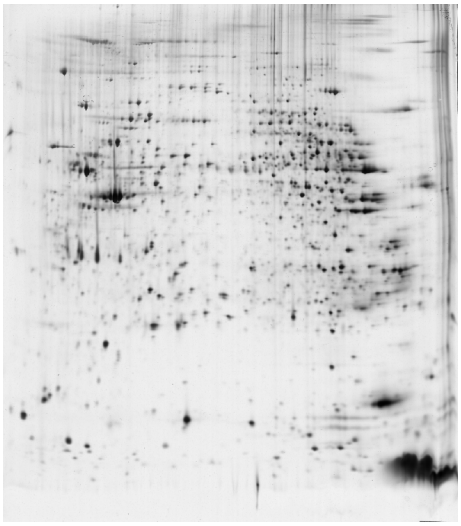


Figure 3.3: *Image of a two-dimensional gel depicting the proteome of human kidney cells.* Dark spots represent accumulations of stained proteins. © Swiss Institute of Bioinformatics, Geneva, Switzerland

An alternative technique for the separation of mixtures of molecules consists in *chromatography*, a procedure which is widely applied in laboratories [236, Chapter 2]. The mixture is dissolved in a solvent yielding a mobile phase. The mobile phase moves over a fixed substance called the stationary phase. As the affinity of the mixture's molecules with the stationary phase varies, they pass it with different velocities. Molecules with a higher affinity for the stationary phase associate with it and move slower, whereas molecules remaining in the solvent pass it faster.

There are several kinds of chromatography methods, for instance paper chromatography, liquid chromatography and gas chromatography. Liquid chromatography, in particular, is frequently utilised in a proteomics context [155]. Compared to 2DGE, chromatography enables the separation based on additional features like the affinity for specific ligands and the ability to solve in water. Its usage has several advantages with respect to 2DGE. It is able to process large sample volumes, is more sensitive, requires no staining, and it can be directly used in conjunction with mass spectrometry, a method to identify proteins (see Section 3.2.2). However, the visualisation of the results is more difficult in comparison to gels. Furthermore, liquid chromatography is less easy to parallelise.

Unfortunately, proteins are not equally well separable. The amount of detectable proteins is assumed to comprise only about 20% to 30% of a cell's proteome according to the Italian proteomics researchers Pier Giorgio Righetti and Egisto Boschetti [179].

> *"Modern proteome analysis is a very complex 'detective story', which might baffle even the most famous investigator, Sherlock Holmes [55]. The reason is that, in any proteome, a few proteins dominate the landscape and often obliterate the signal of the rare ones, so that, when the police reach the scene of the crime, the thin thread of evidence remains hidden."*
>
> Pier Giorgio Righetti and Egisto Boschetti (2007) [179, page 897]

Therefore, techniques had to be developed which enable the analysis of less abundant proteins. Righetti and Boschetti introduced a method called *protein equaliser*. It is based on microscopic beads; that is, small spheres. To these spheres, very short peptides are attached. They may encompass not more than six amino acids. In principle, all different combinations of amino acids of a certain length are applied. Hence, each protein should be able to interact with or rather adsorb to

a bead carrying a suitable peptide. While rare proteins are able to completely adsorb to the corresponding beads, the beads of abundant proteins are saturated quickly. So the remaining unbound proteins can be depleted. The resulting protein distribution is considerably more uniform than the original one. Nevertheless, several proteins are able to bind to more than one peptide yielding to differences in the relative protein abundances.

3.2.2 Protein Identification

After they have been separated, the proteins are to be identified. This can be achieved by using probes such as antibodies, which recognise structures specific for a particular protein or shared by a complete protein class [236, Chapter 3]. Individual proteins can be identified by a technique called *immunoblot* or *western blot* [133, Chapter 24]. Here, proteins previously separated by gel electrophoresis are transferred onto a membrane. Then, antibodies against specific proteins are added. An element of the resulting matrix corresponds to the amount of protein a particular antibody binds with. In order to visualise them, these antibody-protein complexes are stained, finally.

Alternatively, *analytical protein chips*, also called *protein microarrays*, are applicable [236, Chapter 9]. They constitute small slides containing an array of different probes specific for up to several hundred targets. After flooding a chip with analyte, the targets bind with the probes. Then, the protein chip is washed so as to remove unbound proteins. Proteins bound to the chip may be detected by techniques such as fluorescence labelling which allow for a direct visualisation. For the analysis of a cell's complete proteome the number of required probes would equal the number of expressed proteins reaching numbers higher than one million. This problem has been too complex for techniques like immunoblots and protein chips up to now.

Alternatively, a protein's sequence can be determined in order to identify it. The most common method consists in *Edman degradation* named after its Swedish-born inventor Pehr Edman who developed it in the 1960s [60][236, Chapter 3]. The Edman degradation successively removes single amino acids from one side of an amino acid chain. These acids are then analysed by means of chromatography. The maximal efficiency for a single step is approximately 98%, which seems to be very high at first glance. But unsuccessful cuts affect the result of following steps. So, after 35 cycles only 50% of the cleaved amino acids originate from the right position. Therefore, Edman degradation is usually not applied to more than 50 successive amino acids in a single run. Larger chains are broken at specific positions yielding small fragments called *peptides*. The sequenced fragments must then be assembled properly, for example, by analysing the corresponding genetic code. Furthermore, Edman degradation requires large amounts of purified proteins and is only applicable if an amino acid chain's terminus has not been altered, for instance by post-translational modifications [137].

Due to the negative aspects of Edman degradation, novel methods for the determination of amino acid sequences were required. *Mass spectrometry* provided the solution to this problem [137][236, Chapter 3]. In principle, mass spectrometry enables the measurement of ions' mass-to-charge ratios in a vacuum. These ratios are usable for deriving molecular masses.

In order to obtain ions, the substance to be analysed is usually broken down into smaller components. Large polypeptides are cut at certain positions so as to yield smaller peptides, which are ionised. The composition of the original polypeptides is deduced from the resulting mass-to-charge ratios (see Figure 3.4). Besides the determination of previously unknown amino acid sequences, mass spectrometry enables the identification of proteins by comparing spectra or derived sequence

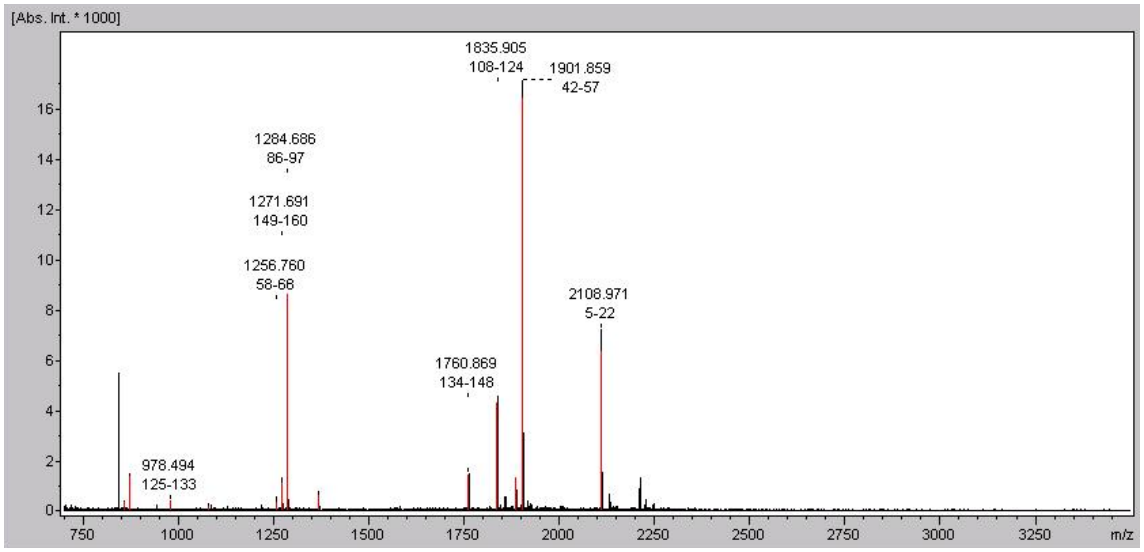


Figure 3.4: *Mass spectrum of a protein originating from Xanthomonas campestris*. The peaks indicate the relative abundances of peptides which possess specific mass-to-charge ratios (m/z). The knowledge of these sequence fragments enables conclusions on the whole polypeptide sequence to be drawn.

fragments with databases.

In the context of proteomics, two techniques are frequently applied in order to ionise the peptides: *matrix-assisted laser desorption/ionisation* (MALDI) and *electrospray ionisation* (ESI). The advantage of these methods, in comparison to alternative mass spectrometry techniques, lies in the fact that peptides under consideration are not damaged.

The generated ions are accelerated by an electric field in the direction of an analyser separating them according to their mass-to-charge ratios. With MALDI frequently the *time of flight* (TOF) analyser, which exploits the fact that heavier ions move slower, is employed. Alternative techniques are, for example, ion traps and quadrupoles. After the peptides have been separated, a detector counts the impacts caused by the corresponding ions. Here, an additional fragmentation of the peptides into smaller ions, for instance by *collision-induced dissociation* (CID) might be useful.

3.2.3 Solving a Protein's Three-Dimensional Structure

A protein's structure directly determines its function [236, Chapter 5]. As a result, proteins with similar structures perform similar functions [236, Chapter 6]. So, the protein function can be predicted based on three-dimensional structures of proteins. Here, even features that are not identifiable by means of the amino acid sequences might be revealed. In order to acquire a sufficient amount of structure data, high-throughput approaches are required. The development of such methods is the goal of *structural proteomics*.

Unfortunately, it is significantly more difficult to obtain structure information than sequence data. Furthermore, so far it is not possible to predict a protein's three-dimensional structure based solely on its sequence [236, Chapter 6]. Therefore, protein structures must be solved directly. Here, two techniques are usually applied: *X-ray crystallography* (XRC) and *nuclear magnetic resonance* (NMR) *spectroscopy*.

X-ray crystallography is a laborious technique employing X-rays, a type of high-energetic elec-

tromagnetic radiation (see Section 3.1), to determine the positions of a protein's atoms [60][236, Chapter 6]. Here, the property that X-rays are diffracted by a molecule's electrons is exploited. Unfortunately, the diffraction caused by a single molecule is too weak. In order to amplify it, a large number of proteins is arranged in a regular lattice or rather a crystal. The creation of such a crystal is a major obstacle for realising high-throughput X-ray crystallography. Nevertheless it can by alleviated by the usage of automated crystallisation workstations enabling a fast optimisation of the crystallisation conditions.

After a crystal has been formed, it is exposed to X-rays and the diffraction patterns are recorded. These patterns are applied to compute an electron density map. Eventually, the structural model is derived. Since several atoms such as carbon, oxygen and nitrogen cannot be distinguished by means of the electron density map, additional information is required. It is usually provided by the amino acid sequence, which has to be known in advance. Furthermore, hydrogen atoms are not detected by this technique.

Nuclear magnetic resonance spectroscopy exploits magnetic properties occurring in several atomic nuclei consisting of an uneven number of protons, an uneven number of neutrons or both. The underlying processes were first reported by the American physicist Edward Mills Purcell and his colleagues in 1946 [169]. In the context of protein structure analysis, common hydrogen (^{1}H) is frequently applied [60][236, Chapter 6]. The nucleus of these atoms rotates causing a magnetic moment, as it is electrically charged. If exposed to a steady magnetic field they can have one of two orientations: parallel or antiparallel to the field's direction. The antiparallel orientation corresponds to a higher energy level than the parallel one. By means of a type of low-energetic electromagnetic radiation, the radio waves, the nuclei can be excited in order to induce a change into the antiparallel orientation. When they return to their original orientation, they emit radiation which can be measured.

In contrast to XRC, NMR spectroscopy does not necessitate a crystal. Instead, it applies proteins in solution. The determination of a protein's structure becomes possible, as electrons in bonded atoms influence the frequency of the radio waves, which cause a flip to the antiparallel direction. Hence, the chemical environment of a nucleus affects the result and yields a number of distance constraints corresponding to pairs of interacting atoms. Provided the number of obtained constraints is high enough, models can be deduced [236, Chapter 6]. These models include the positions of numerous hydrogen atoms, which are not provided by XRC.

3.2.4 Protein Quantification

Besides a protein's structure and sequence, its abundance might be of interest. So, for example, proteomic changes between different cell types, states or caused by varying environmental conditions are revealed [236, Chapter 4]. Unfortunately, these differences are rather small resulting in difficulties to measure them.

The approaches to protein quantification can be divided into two principal categories: methods for examining particular proteins in complex mixtures and more general techniques measuring the total protein abundance. Specific proteins can be quantitated by means of immunoblots or protein chips (cf. Section 3.2.2), as these techniques, besides the identification of proteins, enable the abundance of specific proteins to be measured. The total amount of protein can be determined by measuring the absorbance of ultraviolet light. Here, the proteins are not damaged and can be applied for further investigations. Alternatively, various staining approaches may be utilised.

In order to investigate the changes of the abundance of large numbers of proteins in parallel, two-dimensional gel electrophoresis as well as a combination of liquid chromatography and mass spectrometry are employed. In the case of 2DGE, digital images of the gels under consideration have to be acquired first. Then, corresponding spots of protein accumulations must be detected. Finally, features such as a spot's size or density can be evaluated. The registration of gels can be avoided by applying multiple dyes to different groups of proteins simultaneously.

In contrast to 2DGE, mass spectrometry (cf. Section 3.2.2) is based on peptides obtained by cleavage of the amino acid chain at particular positions rather than complete proteins. It is usually applied in conjunction with separation methods such as chromatography. In order to allow for a quantitation, the proteins or rather their peptides are tagged by atoms of a specific element with different masses, which are called *isotopes* [60]. The isotopes cause shifted peaks in the mass spectrum. Hence, conclusions about the differences in proteins' abundances can be drawn by comparing the peaks' heights.

Although the labelling of proteins might be realised in living cells, that is *in vivo*, the cells must be destroyed before the actual identification or quantification of proteins, by both, 2DGE and mass spectrometry. So, it is difficult to examine dynamic processes taking place in intact cells.

3.3 Computational Techniques

The experimental proteomic techniques have yielded vast amounts of data on proteins, especially their sequences and structures, which are stored in databases [236, Chapter 5]. In addition, amino acid sequences have been derived from nucleotide sequences resulting in even more available information about proteins. So, it is not surprising that the analysis of data contained in such databases constitutes an essential part of proteomic research.

For database searches, it is usually assumed that proteins performing similar functions have similar three-dimensional structures, since the structures dictate how proteins interact with other molecules. This structure in turn is determined by the amino acid sequence. So, there exists an intimate relationship between a protein's function and its amino acid sequences and structure, respectively.

Similarities between proteins are not necessarily caused randomly. According to the theory of evolution, which is discussed in Section 3.3.1, they might be a result of evolutionary processes; that is, they may originate from common ancestors. Section 3.3.2 details the usage of sequence information for inferring proteins' functions and principal problems of this method.

Nevertheless, proteins fulfilling equivalent tasks do not need to be evolutionarily related. They may only share some very special features which are crucial for their function. Therefore, methods have been developed analysing such common features (see Section 3.3.3). Furthermore, a proteins three-dimensional structure, if available, can provide information concerning a protein's function (see Section 3.3.4). Unfortunately, this approach is very computationally expensive. So, it is beneficial to incorporate prior knowledge reducing the set of structures to be compared.

Finally, conclusions about possible interactions and locations of proteins can be drawn based on available databases. Such methods are introduced in Sections 3.3.5 and 3.3.6. But the results yield only hints; they should be verified by experimental techniques (cf. Sections 3.4 and 3.5).

In order to assess the quality of results obtained from database searches, the quality of the applied data must be taken into account, as the databases contain errors [236, Chapter 5]. If such

incorrect information is used to derive novel data, the errors propagate. Moreover, the incorrect usage of available tools might introduce new problems. So, computational methods should be applied carefully.

3.3.1 Exploiting Evolutionary Relations

At the beginning of the 19th century, the theory of *evolution* was developed. Although Charles Darwin is frequently considered its founder, several historical personages were involved in the discussion on it how it works and what implications result from it [48, Historical Sketch] [108]. The French naturalist Jean-Baptiste Lamarck, one of the early pioneers in the field of evolution, argued that all species have evolved from other species. He suggested that changes depend on the degree of usage of the respective trait, which are passed to the next generation.

> *"On conçoit de là qu'un changement de circonstances forçant les individus d'une race d'animaux à changer leurs habitudes, les organes moins employés dépérissent peu à peu, tandis que ceux qui le sont davantage, se développent mieux et acquièrent une vigueur et des dimensions proportionnelles à l'emploi que ces individus en font habituellement."*[2]
>
> Jean-Baptiste Lamarck (1809) [125, page v]

Besides many others, the well-known German poet and scientist Johann Wolfgang Goethe contributed ideas to the debate on the theory of evolution [108]. But until 1859 when Charles Darwin published the first edition of his famous book entitled "On The Origin of Species by means of Natural Selection; or, the Preservation of Favoured Races in the Struggle for Life", evolution was largely considered pure speculation.

In contrast to Lamarck, who regarded the use and disuse of organs as the main force of evolution, Darwin attributed evolutionary modifications largely to natural selection, a principle, which is mainly based on the struggle for life.

> *"Owing to this struggle, variations, however slight and from whatever cause proceeding, if they be in any degree profitable to the individuals of a species, in their infinitely complex relations to other organic beings and to their physical conditions of life, will tend to the preservation of such individuals, and will generally be inherited by the offspring. The offspring, also, will thus have a better chance of surviving, for, of the many individuals of any species which are periodically born, but a small number can survive. I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection, in order to mark its relation to man's power of selection."*
>
> Charles Darwin (1872) [48, page 45]

The source of the individuals' variations remained unclear and a topic of intense dispute, since its molecular causes were unknown. At the same time, the Austrian Augustinian abbot Gregor Johann Mendel discovered his *laws of inheritance* by investigating pea plants. He assumed that

[2] Translation: "One conceives that where a change of circumstances forces the individuals of a race of animals to change their habits, the organs which are less employed shrivel gradually, whereas the ones that are applied more develop better and gain a vigour and dimensions proportional to the usual usage by these animals."

the hereditary information of living organisms is composed from factors which represent traits. In terms of modern language, these factors roughly correspond to genes. His laws describe how these factors are transmitted to a new generation of individuals. In addition, he concluded that all factors occur twice. This insight accords with our knowledge about *diploid* cells, which have two copies of each chromosome.

Mendel's laws and Darwin's evolutionary theory were considered incompatible [191]. In 1911, the group of the American geneticist Thomas Hunt Morgan started to fuse both theories. They investigated the inheritance of traits by means of the fruit fly *Drosophila melanogaster*. With this, two important means of evolution were discovered: *mutation* and *recombination*. The recombination of parental genes explains Mendel's laws, while mutation ensures the variability required by natural selection. The origin of these mechanisms has later been elucidated by molecular biology.

In spite of these developments, a new theistic theory called *intelligent design*, which aims at replacing Darwin's evolutionary theory, has emerged during the last few decades [20]. The number of its proponents is rising noticeably, especially in the United States of America. They are exploiting gaps in our current scientific knowledge, which they are trying to fill by claiming that the course of evolution has been controlled by supernatural entities. However, intelligent design has not provided any proof sufficient to refute the theory of evolution by natural selection up to now. In fact, current research such as the comparison of sequenced genomes of several organism rather provides indications in favour of natural evolution, since there appear to be close relationships between very different species such as the fruit fly and yeast [220].

The knowledge that all living organisms are products of evolution can be applied to derive information on an unknown protein's function if several evolutionarily related proteins have already been examined. Only functional proteins derived from functional ancestors occur in nature. So, if a protein's function has been conserved for numerous generations, amino acids which are required for performing this function as well as their relative positions in the protein's three-dimensional structure must have been conserved, too.

3.3.2 Alignment of Amino Acid Sequences

The function of proteins can be inferred from evolutionarily related, i.e. *homologous* proteins, by comparing the sequences of their amino acid chains. Functional segments of a polypeptide's amino acid sequence can be determined by identifying evolutionary relationships between corresponding proteins first and looking for conserved regions afterwards. If the function of one of these proteins is known and has been associated with particular segments of the corresponding amino acid sequences, conclusions about the function of the second protein can be drawn if the functional residues have been conserved.

Instead of solving both problems separately, evolutionary relationship is usually assumed [225, Problem 1]. Then, conserved amino acids can be determined in such a way as to maximise the probability that both proteins originate from a common ancestor. Finally, the corresponding probability is computed. So, evolutionary relationships might be revealed. Unfortunately, even related proteins might have only a low fraction of amino acids in common, for example, the sequence identity of the fluorescent proteins *Aequorea* GFP and DsRed amounts to less then 30% (cf. Section 3.5.1) [140]. Furthermore, detected similarities do not guarantee that two proteins perform the same function, even if the proteins are evolutionarily related (see Figure 3.5). Numerous proteins showing strong sequence similarities but performing different functions are known [236,

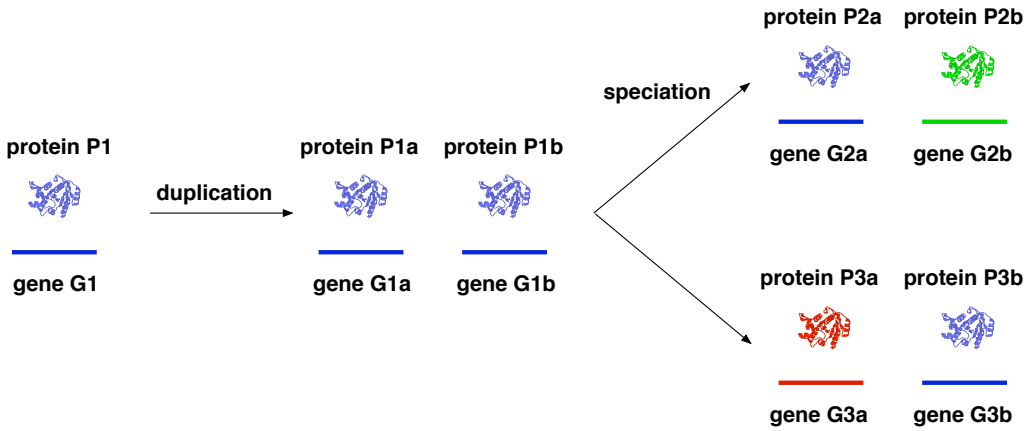


Figure 3.5: *Evolution of proteins with different functions from a common ancestor.* Gene G1 of a species of organisms encodes for a protein P1. This gene is duplicated as the species evolves. Then, two populations are separated by environmental changes. They become independent, which results in two new species. Since only one gene is required which encodes for a protein with the function of the original protein P1, the additional copies might develop a new function. So, although P2a and P3a are evolutionarily related, their functions need not be the same.

Chapter 5]. In contrast, splice variants can be distinguished from homologous proteins with high accuracy [208].

The function of proteins having a quaternary structure, i.e. that they are constructed from at least two polypeptides, is more difficult to analyse. Such proteins might have different functions, even though segments of several amino acid chains are very similar. Moreover, effects resulting from the higher-level structure of a protein cannot be examined directly.

On the whole, the accuracy of methods exclusively analysing the amino acid sequences in order to determine the function of proteins is limited. Nevertheless, the alignment of such sequences constitutes a very important method for the determination of the respective proteins' functions, since biochemical experiments are avoided resulting in a significantly higher number of investigations, which can be performed.

The foundation of sequence alignment algorithms was laid by Saul B. Needleman and Christian D. Wunsch in Chigago in 1970. Their algorithm allows for the global alignment[3] of two amino acid sequences using iterative matrix computations [152]. Temple F. Smith and Michael S. Waterman modified this algorithm in order to enable the determination of local alignments[4] in 1981 [205].

Unfortunately, it is not possible to extend these approaches to cope with multiple sequences, since the computational effort would be too high. Nevertheless, such multiple alignments are of interest to biologists, as they can reveal evolutionary and structural relationships between different species. Thus, special approaches have been introduced, which exploit similarities in order to determine an order of binary matches [225, Problem 1][38].

A multiple sequence alignment can be represented by means of a *profile*. Here, the probability that an amino acid occurs at a specific position is approximated. Then, additional sequences can be aligned with the profile. Another popular technique consists in the application of *hidden Markov models* [84][166][172]. So, models of particular classes of proteins can be derived automatically enabling a discrimination from other types of proteins.

The data required for performing alignments can be obtained from *UniProt* – the universal pro-

[3] alignment across the complete sequence
[4] alignments of certain regions

tein resource [221]. It gives access to a huge amount of amino acid sequences including functional annotations. The UniProt archive contains more than 14.8 million sequences[5] up to now.

3.3.3 Feature-Based Analysis of Proteins

Besides investigating evolutionary relationships, patterns derived from groups of proteins exhibiting particular features can provide additional information about a protein whose function is unknown. So, for example, a common sequence pattern shared by proteins responsible for the same function can be applied to other proteins. In contrast to evolutionary relationships leading to large conserved segments of a polypeptide chain, functional similarities might be represented by a small number of amino acids [225, Problem 2]. These amino acids usually require a specific spatial relation relative to each other, which is ensured by the remaining elements of the sequence.

Unfortunately, many proteins whose sequences are stored in protein databases are only partially or even not associated with a function that has been experimentally verified. So, it is difficult to obtain training samples, which are required for deducing rules or common patterns. A collection of such patterns which might be defined as regular expressions [204, Chapter 8] or profiles is obtainable from the PROSITE database [97].

3.3.4 Examining Structural Similarities

As detailed in Section 3.2.3, knowledge on proteins of unknown function cannot only be deduced from their amino acid sequence. If known, a protein's three-dimensional structure might provide crucial information about its function, as well [225, Problem 6]. It even enables the detection of relationships that are not recognisable by sequence-based methods, since structure is considerably better conserved than sequence [236, Chapter 6]. By superposing the structure of a protein under analysis with the one of a protein whose function is known, similarities may be revealed. Here, even local similarities can indicate a specific function.

Nevertheless, like proteins with similar amino acid sequences, proteins with similar structures may perform different functions (cf. Section 3.3.2) [236, Chapter 6]. Since based on protein structure more distant evolutionary relationships are revealed, it is assumed that the functional diversification is even greater. In addition, similar structures performing the same function might have evolved independently. Such proteins which are not evolutionarily related are termed *analogous proteins*.

Among others, the *SCOP database* provides an ordering of proteins with respect to their structure. The structures themselves can be accessed via the *protein data bank* (PDB) [14], which contains information on 47,509 macromolecules[6] including RNA and DNA.

Unfortunately, it is not possible to perform a superposition with all known protein structures, as this method is extremely computationally expensive. Therefore, prior knowledge, for instance about evolutionary relations and chemicophysical properties, must be incorporated.

3.3.5 Predicting Protein Interactions

Numerous biological functions are based on the interaction of proteins [225, Problem 7], for example by forming complexes or regulating the expression. But proteins do not only interact with

[5] UniProt Release 12.2

[6] version released at the 27th of November 2007

other proteins: they can bind to small molecules termed *ligands*, as well [225, Problem 8]. Such ligands are capable of changing the structure of a macromolecule; for instance hemoglobin alters its structure if oxygen atoms bind. Furthermore, interactions with macromolecules other than proteins might occur, for example with nucleic acids. The diverse types of interaction have usually been treated differently.

As the number of known protein structures is much smaller than the number of available amino acid sequences, it would be beneficial to predict interactions based on these sequences. This can be accomplished by regarding different species. For example, in the case that two proteins occur separately in one species and form a complex in another species, they are likely to interact. The co-occurrence of two proteins in a large number of species – that is, both of them are either present or absent – might indicate their interaction as well. Another possibility to determine likely interactions consists in examining the rate of evolution of two proteins in question. Since their interaction would imply coevolution, they presumably do not interact if they evolve with different rates.

Unfortunately, these methods do not create a realistic image of true protein interactions, as they might be ambiguous and hard to understand [225, Problem 7]. Therefore, experimental validation is mandatory. Moreover, due to the large amount of data, very fast methods that enable high-throughput processing are required (see Sections 3.4 and 3.5).

In contrast to sequence-based techniques, structure-based methods try to position two proteins in question in such a way that they form a complex. Therefore, this method is called *protein–protein docking* Unfortunately, the structure of proteins varies depending on whether they are bound or unbound. So, besides identifying complementary surfaces, the flexibility of the structure must be taken into account as postulated by the American cell biologist Daniel E. Koshland in 1958 [122].

In order to dock them, the proteins are usually represented as three-dimensional objects, for example by means of surfaces, spheres or voxels[7] in a three-dimensional grid [225, Problem 7][279]. Then, patches of their surfaces are compared. This procedure is usually very computationally intensive. As a result, several possible solutions are available which need to be ranked by a scoring function to determine the best one. The choice of the scoring functions therefore constitutes a critical part of the docking system.

The goal of *protein–ligand docking* consists in the determination of binding sites, positions and affinities. Hence, knowledge about promising modifications of a molecule leading to an increase in its affinity or specificity might be obtained. A comprehensive overview on available methods is given in [225, Problem 8].

3.3.6 Predicting Protein Locations

Based on knowledge of a protein's amino acid sequences, its subcellular location is predictable. For this classification task, features such as targeting signals (see Section 2.4), which control the transportation, specific structural elements as well as information on homologous proteins have been exploited. The results are stored in databases, e.g. PSORTdb [178], together with experimentally obtained data. To date[8], PSORTdb encompasses 2,171 experimentally determined as well as 922,714 computationally derived entries.

[7] three-dimensional pixels
[8] 28th December 2007

Owing to their dissimilar structure (cf. Section 2.1.1), different approaches are required for prokaryotic and eukaryotic cells. Eukaryotic cells contain a higher number of compartments which have to be distinguished. Paul Horton and his co-researchers [88] employed 14 classes, four of which represented dual localisations. The performance was rather low. With respect to Fungi, the accuracy reached values higher than 50% only for five classes; proteins of five other compartments were completely misclassified. Alternative approaches which employ a limited set of possible locations achieve considerably better results [163]. Nevertheless, they are clearly outperformed by experimental methods (cf. Section 3.5) [29].

In Bacteria, usually not more than five different compartments are considered [71][15]. This number may be increased by additional dual localisations, as well. Unlike in the case of Eukaryotes, here, the small number of compartments can be justified with the simple structure of prokaryotes.

3.4 Protein Interactions

As discussed in Section 3.3.5, the computational methods for predicting protein interactions based on sequence or structure information are not sufficiently reliable. Hence, there exists a great variety of experimental procedures for examining the interactions of proteins. The branch of proteomics that is devoted to this type of research is called *interaction proteomics*.

In principle, protein interactions can be examined using XRC or NMR spectroscopy. Both of these methods are not only suited for investigating protein–protein interactions but other interaction types, as well. They yield very detailed information but have not been amenable to high-throughput application, yet.

Protein interactions are more efficiently detectable using a high-throughput technology called the *yeast two-hybrid system* (cf. Section 3.4.1). It enables the analysis of binary protein-protein interactions; i.e., direct interactions of two proteins. Nevertheless, more complex interactions might be revealed by performing investigations on a large scale and constructing interaction maps of the complete proteome [102]. These interaction maps visualise proteins and protein complexes as nodes. Links between such nodes correspond to interactions.

But the yeast two-hybrid system is not the only technology enabling an efficient analysis of protein interactions. In Section 3.4.2, a selection of alternative methods is introduced. Here, the focus is on potential and existing high-throughput technologies.

3.4.1 The Yeast Two-Hybrid System

The yeast two-hybrid system constitutes a crucial technology enabling the global analysis of protein–protein interactions. It exploits special properties of the reproduction cycle of budding yeast as well as features of special proteins that control transcription in eukaryotic cells and are thus called *transcription factors*.

Budding yeast, like all eukaryotic cells, occurs in two forms: a haploid and a diploid one [133, Chapter 14]. *Haploid* cells contain one copy of each chromosome, whereas *diploid* cells possess two copies. So, in the case of budding yeast, haploid cells encompass 16 and diploid cells 32 chromosomes. Each of these types can reproduce asexually in a process referred to as *mitosis*. Here, the number of chromosomes present is doubled and afterwards corresponding chromosomes are equally distributed between two daughter cells. In contrast, the number of chromosomes is

halved by *meiosis*, in order to obtain gametes for sexual reproduction. Here, like in mitosis, the number of chromosomes is doubled. But then four haploid gametes are formed. The result of a fusion of two of these gametes consists in a diploid cell.

In order to perform yeast two-hybrid experiments, a transcription factor is divided into two fragments: the DNA-binding domain and the activation domain (see Figure 3.6). Then fusion proteins are constructed. The bait protein, whose interactions are to be analysed, is combined with the transcription factor's DNA-binding domain. The corresponding DNA is introduced into a haploid yeast strain and expressed. These proteins are not able to activate transcription on their own.

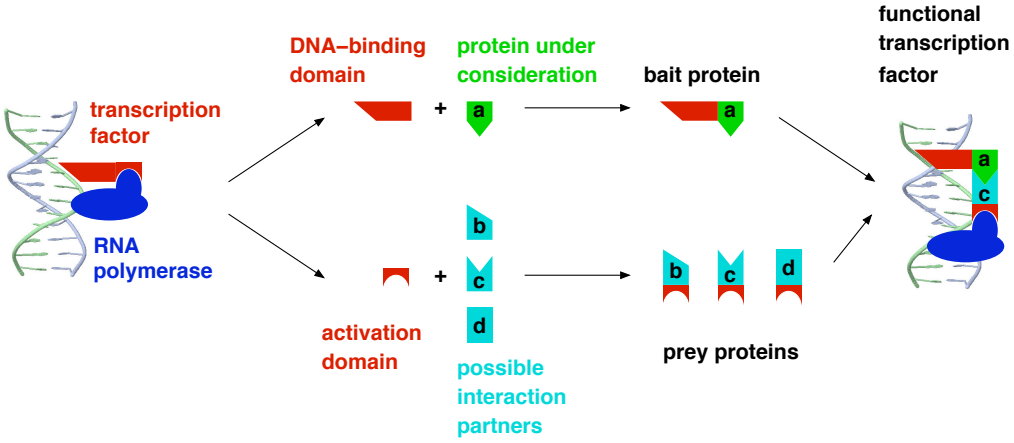


Figure 3.6: *Principles of the yeast two-hybrid system. The protein whose interactions are to be analysed (a) as well as possible interaction partners (b, c, d) are fused with fragments of a transcription factor – the DNA-binding domain and the activation domain. Provided two proteins interact, a complex encompassing both domains is formed, which enables the transcription of a reporter gene. So, the interaction can be visualised..*

Besides the bait protein, prey proteins are constructed from possible interaction candidates by fusing them with the transcription factor's activation domain. In principle, all other proteins are applicable as prey. The respective haploid yeast strains are pair-wisely mated with the bait strain. Assuming the bait protein interacts with a prey protein, functional transcription factors are created, as the activation and DNA-binding domain are very close to each other. It can be detected by including a *reporter gene* whose transcription is controlled by the applied transcription factor. If expressed, this reporter gene results in an observable change of the yeast's appearance. Examples for reporter genes and their products are introduced in Section 3.5.1.

Using the yeast two-hybrid technique tens of thousands of protein interactions can be systematically scanned in a living system by performing pair-wise matings of haploid yeast strains [236, Chapter 7]. As these protein interactions are the basis of numerous biological processes, the functions of a large number of proteins can be inferred from interactions they are involved in [56]. So, the yeast two-hybrid is a very popular method for the determination of protein functions.

Nevertheless, the yeast two-hybrid system has several important constraints. Since fusion proteins need to be constructed for every protein being part of the analysis, comprehensive knowledge on the considered organism's genome such as its complete nucleotide sequence is required. Furthermore, the quality of the results is relatively low [56][236, Chapter 7]. Firstly, similar large-scale studies yielded different results. Only 10% to 20% of the detected interactions in one study occurred in the second one (see [101] and [238], for example). This indicates that the investigations

are either not comprehensive or depend strongly on the experimental conditions. Secondly, large-scale variants of the yeast two-hybrid system are assumed to miss numerous interactions, which might, for instance, result from the preparation of the fusion proteins. Another reason consists in the formation of the transcription factor in the yeast's nucleus; proteins which do not normally localise there could behave differently or post-translational modifications might not happen as in their natural environment. Thirdly, 10% of the occurring proteins contain a transcription activator themselves. Thus, even if no interaction takes place, the reporter gene is transcribed leading to false positives. In addition, proteins which might be usually separated are brought together causing unnatural interactions. Eventually, prey and bait proteins might interact indirectly involving a third protein linking to both of them [102].

Due to the addressed problems a verification of the results obtained by means of the yeast two-hybrid system is necessitated. This can be achieved by comparison with reliable interactions or looking for overlaps with alternative experiments such as subcellular localisations [102] (cf. Section 3.5).

3.4.2 Alternative Approaches to Protein Interaction Analysis

Besides utilising the yeast two-hybrid system, affinity-based methods such as *affinity chromatography* can be employed [236, Chapter 7]. Here one protein or another possibly interacting molecule is immobilised functioning as stationary phase. Then, cell lysate is passed over it. Proteins interacting with the stationary phase adsorb to it, while other proteins just pass it. Retained proteins can than be identified by means of techniques such as mass spectrometry and immunoblotting (cf. Section 3.2.2). In particular, mass spectrometry has proven beneficial in order to enable large-scale screens. Unfortunately, such experiments are supposed to miss the majority (up to 60%) of the occurring interactions. Nevertheless, similar principles are applied in alternative approaches such as phage display and functional protein chips.

Phage display introduces relevant DNA sequences into *bactriophages* [133, Chapter 9], i.e. viruses infecting prokaryotes, where it is expressed as coat protein [47][236, Chapter 7]. In order to analyse interactions of the complete proteome, a phage library is formed containing phages displaying every occurring protein. Then slides are coated with proteins under consideration and the phage library is pipetted on them. If two proteins interact, the phage binds to the slide. Otherwise it is washed away. Finally, phages displaying interacting proteins can be separated from the slides and used to infect bacteria. So, the interacting protein's DNA is amplified considerably enabling the nucleotide sequence to be determined and compared to databases.

In principle, phage display can be completely automated and used for interaction analysis on a proteomic scale [236, Chapter 7]. However, such large-scale screens are not common. Instead, phage display is frequently applied in conjunction with alternative experimental methods such as the yeast two-hybrid system, which enables reliable results. Furthermore, it is employed for the determination of specific proteins linked with diseases and are therefore referred to as diagnostic or therapeutic *targets* [47].

Functional protein chips allow for a systematic analysis of various interaction types. The proteins that are to be analysed are attached to the chip at different positions on a grid. After the chip has been exposed to the analyte, interactions can be rendered visible based on a label of the interaction partner or by means of label-free techniques such as *surface plasmon resonance spectroscopy*, which exploits optical properties of the surfaces of gold-coated glass chips [236, Chapter 7].

The usage of functional protein chips has several advantages. Firstly, the required sample volumes are very small and large numbers of proteins can be analysed in parallel. Secondly, the experiments are amenable to automation enabling high-throughput investigations of protein functions if a sufficient amount of knowledge about the organism's genome is available. Moreover, the results have been reported to show a high reciprocity [156].

On the other hand, today's functional protein chips are restricted to simple organisms like yeast whose proteome comprises about 6,000 proteins [238]. Chips for higher eukaryotes are significantly more difficult to develop, since the influence of effects such as alternative splicing and post-translational modification is increasing; for example, the human proteome is assumed to consist of more than one million proteins. Therefore, protein chips have not been well-suited for the analysis of more complex eukaryotic proteoms, up to now. In addition, several proteins require differing environmental conditions, which cannot be realised on a single chip. Thus, the results do not reflect real biological processes. Eventually, the location of proteins is neglected. So, proteins usually occurring in different subcellular compartments may interact.

3.5 Location Proteomics

In addition to the techniques discussed above, the localisation of proteins in living cells is able to reveal important information on the functions and interactions of proteins. Provided two proteins occur in the same subcellular compartment of a cell at the same time, they might interact. This knowledge is applicable for the simulation of cell behaviour which might facilitate the investigation of diseases as well as the development of innovative drugs and vaccines [29][178]. So, location proteomics could contribute significantly to current proteomic research. Nevertheless, little attention has been paid to it compared to other branches of proteomics [145].

Besides the investigation of co-localisations, conclusions about a protein's function can be drawn solely based on its own location. So, for example, proteins occurring in the mitochondria are likely to be involved in cellular respiration, whereas proteins localised in the lysosomes might take part in the process of cellular digestion (see Section 2.1.1).

A common approach of determining the subcellular location of proteins consists in the examination of fluorescence microscope images [32][29][96][127][147], which is especially well-suited for the analysis of intact cells. In these intact cells, proteins can be observed in their natural environment. In contrast to other proteomic methods such as the yeast two-hybrid system (cf. Section 3.4.1), cells of higher eukaryotes including humans can be applied easily. Therefore, processes such as alternative splicing and post-translational modifications are accounted for.

In order to localise proteins, they are frequently fused with a fluorescent protein, for instance with the *green fluorescent protein* or one of its spectral variants [235] (see Section 3.5.1). So they become visible under a fluorescence microscope (see Section 4.2.6). Reference patterns for specific subcellular locations can be obtained by using electron microscopy (see Section 4.4) and proteins with well-known locations [203], or performing counterstaining with exogenously added dyes [29][90], for example *4',6-diamidino-2-phenylindole* (DAPI) [262] and *chloromethyltetramethylrosamine* (MitoTracker Orange) [132]. In conjunction with living cells, such dyes must be used very carefully, since they might adversely influence the outcome of an experiment or damage the cells [41][196]. In contrast, fluorescent proteins are non-invasive – that is, they do not cause any significant changes to a cell's behaviour [41].

If, besides the subcellular locations, interactions of multiple proteins are to be analysed, additional techniques are required. Such methods are introduced in Section 3.5.2. In contrast to the procedures discussed in Section 3.4, these interactions can be directly observed in living organisms.

Since proteomics aims at analysing the whole proteome, the evaluation of acquired images should be performed automatically. Here, automated image analysis plays a key role. Therefore, the goal of my thesis consisted in the development of such methods. Several existing approaches are introduced in Section 3.5.3.

3.5.1 Fluorescent Proteins

Various biological processes can be visualised by means of fluorescent proteins. These proteins are able to realise a type of *luminescence*. Luminescence describes the emission of photons caused by the excitation of electrons [60]. It can be divided into *fluorescence* and *phosphorescence*. Fluorescence is an immediate reaction to excitation, whereas phosphorescence occurs over a time period longer than 0.0001s. The component of a protein which actually performs fluorescence is termed *fluorophore* [144, Chapter 11].

In order to excite electrons to a higher state, photons of specific wavelengths, defined by the *excitation spectrum*, are required [144, Chapter 11][167, Chapter 5]. When an electron falls back to its original state, a new photon with a lower frequency is emitted, as some energy is transformed, especially into heat. As the amount of transformed energy differs, light of several wavelengths is emitted by a specimen – the *emission spectrum*. The difference between both spectra is called *Stokes shift*, named after the Irish mathematician and physicist George Gabriel Stokes, who investigated the phenomenon of fluorescence in the middle of the 19th century [209]. Large Stokes shifts are desirable, since they enable a better separation of exciting and emitted light.

The most popular fluorescent protein, called simply *green fluorescent protein* (GFP), was first mentioned by the Japanese biochemist Osamu Shimomura from Princeton University and his co-researchers in 1962 [199]. They discovered it by chance during investigations of a bioluminscent protein referred to as *Aequorin* that occurs in the jellyfish *Aequorea aequorea*, also termed *Aequorea victoria*, as well. Similar fluorophores have been observed in other marine species, for instance *Obelia*[9] and *Renilla*[10]. The structure of *Aequorea* GFP is depicted in Figure 3.7.

GFP encompasses 238 residues [157]. Wild-type GFP has two peaks in its excitation spectrum: a major peak at 395nm–397nm and a minor peak with only one third of the major peak's height at 470nm–475nm; the peak of the emission spectrum is at 504nm which corresponds to a green colour [235]. Based on the wild-type GFP, novel fluorescent proteins have been created: the *yellow fluorescent protein* (YFP), the *cyan fluorescent protein* (CFP) and the *blue fluorescent protein* (BFP). These proteins were named after the major colour of their emission spectrum. But not only the emission spectra have been altered: the excitation spectra have been modified as well. As a result, the excitation spectra of some proteins overlap with the excitation spectra of others (see Section 3.5).

In order to investigate several proteins simultaneously, the availability of a great variety of fluorescent proteins with disjoint emission and excitation spectra is required. Unfortunately, until 1998, GFP variants reached only emission peaks at about 518nm–529nm which lies in the yellow

[9] hydroids
[10] see pens

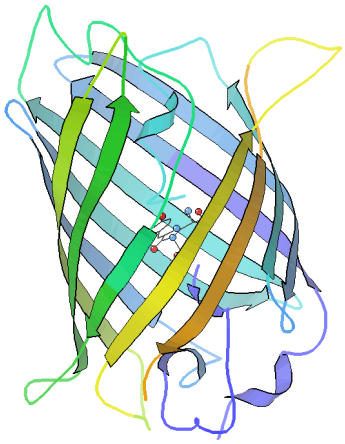


Figure 3.7: *Structure of the green fluorescent protein.* GFP exhibits a barrel-like structure composed of eleven β-strands [157]. Furthermore, an α-helix carrying the fluorophore is situated inside the barrel. The fluorophore has been visualised using a ball and stick model.

spectrum (cf. Section 3.1). As the spectrum of visible light stretches from 400nm to 750nm, more than half of the available range was unused. Therefore, Mikhail V. Matz and his colleagues at the Russian Academy of Science searched for similar proteins occurring in species related to *Aequorea*, in particular reef corals [140]. Remarkably, the examined organisms are not able to perform bioluminescence. Matz and his co-researchers assumed that such fluorescent proteins must be present, as these life forms are likely to share a common ancestor.

The effort paid off. They discovered six fluorescent proteins which are related to *Aequorea* GFP, two of which exhibiting novel spectral characteristics. The length of these proteins varies between 225 and 266 residues. Intriguingly, the sequence identity with respect to GFP is less than 30%, but all elements of secondary structure are observable. One of these six proteins, *DsRed*, which has an emission maximum at 583nm has become popular, in addition to *Aequorea* GFP [11][77][216][264].

Today, a great variety of fluorescent proteins covering a wide range of emission and excitation spectra is available. They mainly originate from *Aequorea* GFP or DsRed. Comprehensive summaries are given in [41] and [198]. In principle, up to four different proteins can be applied in parallel without considerably influencing each other.

In order to determine the location of proteins, fusion proteins are constructed, i.e., a reporter gene encoding for a fluorescent protein is incorporated into the gene corresponding to the protein under consideration. If the resulting gene is expressed, the formed proteins contain a fluorescent component. Nevertheless, they perform their usual function. But now, they are observable. After their expression and excitation, they emit light of a specific wavelength that can be observed through a microscope. The corresponding images show accumulations of the investigated protein and thus enable its localisation.

3.5.2 Confirmation of Proteins Interactions

As a co-localisation is only a hint and no evidence for an interaction, it must be confirmed by additional studies, for example by *fluorescence resonance energy transfer* (FRET). Here, two fluorophores emitting light of different wavelengths, for example CFP and YFP (cf. Section 4.2.6), are conjugated to possibly interacting proteins [180]. In order to perform FRET, the fluorophores are chosen in such a way that the emission spectrum of one (the donor) overlaps with the excitation spectrum of the other (the acceptor). As a result, the acceptor emits light if it is placed in close

proximity to the activated donor, which happens if the proteins interact. This light can be measured and applied to the detection of interactions. Such confirmed interactions are very reliable [236, Chapter 7].

In order to gain independence from the intensity of a fluorescent signal, *fluorescence lifetime imaging microscopy* (FLIM) can be applied [41]. Here, the lifetime of the fluorescence is measured. This limited lifetime is a result of a decay of the fluorescence intensity after excitation. It depends on the employed fluorophores. So, FLIM images visualise differences in fluorescence decay rates.

3.5.3 Image Analysis

The analysis of fluorescence micrographs showing tagged proteins enables their subcellular location to be determined. An introduction of current methods is given in [214]. Even complete proteomes have been localised [96]. The micrographs may be recorded using two or three dimensions, depending on the employed microscopy method (see Chapter 4). Usually magnifications between 40× [44][121] and 100× [16][96] are applied. In principle, here, the highest magnification possible should be preferred, as subcellular structures are to be visualised [89]. Unfortunately, the magnification is limited by the pixel size of the camera, the wavelength of the emitted light and the microscope's properties, in particular, the *numerical aperture*[11] [144, Chapter 6].

Accumulations of fluorescent proteins are depicted as bright spots, which possess a specific size, shape and position within the corresponding cells (cf. Figure 3.8). So, a trained experimenter, who has some knowledge of the cells used, is capable of distinguishing different location patterns [89][96]. Sometimes the investigation is supported by auxiliary microscope images [121] or image editing software [203].

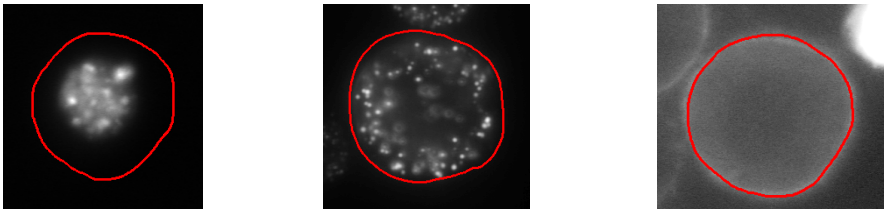


Figure 3.8: *Protein distributions in insect cells (Sf9).* This figure depicts micrographs of tagged proteins that are localised in three different cell compartments. The red contours represent the surrounding cells, which were manually extracted from corresponding bright-field images by biological experts.

To date, cell biologists have frequently evaluated the images they acquire themselves resulting in a limit of the number of images which can be considered as well as a bias with respect to the experimenter's training and experience [89]. So, the usage of automated procedures able to solve these problems would be beneficial. Unfortunately, the examined cells are not necessarily visible in these fluorescence images. As the powerful visual information processing system humans can resort to in order to recognise them nevertheless is not available for automated methods, additional knowledge is required to associate fluorescent spots with specific cells. This knowledge

[11] The numerical aperture depends on the refractive index of the medium between the objective lens and a specimen as well as the angle up to which light from the specimen is accepted.

is frequently obtained by additional stains [44][127] or manually selecting images of single cells [16].

If images of single cells are available, a *classifier* can be trained with the pattern of the subcellular locations they show. These subcellular distribution patterns are often described by means of numerical features [36][89]. As each cell image is associated with a specific location the whole process is termed *supervised learning*. After training has been finished, the classifier is able to classify unknown images, that is, to assign an appropriate location to new images of fluorescent patterns from single cells. Locations which were not used for training cannot be recognised or, even worse, are classified incorrectly.

Based on the classification of numerical features, efforts have been made to automatically extract images depicting tagged proteins from on-line journal articles [91][148]. The goal of this work has consisted in the creation of a common knowledge base collecting images of proteins localised in different organelles and cells.

For special applications necessitating only the investigation of a specific cell compartment a more detailed analysis is performed. In [173], for example, compartments called centrosomes are detected exploiting radial symmetries. So the individual examination of these centrosomes is enabled. But due to the large number and small size of some cell organelles such as the mitochondria, this approach cannot be generalised to all protein localisations. Furthermore, it is not able to distinguish between different locations like the general feature-based method.

In the future, the automated learning of new categories will become more important [36][145] resulting in the need for appropriate unsupervised learning methods and a classifier's ability to detect unknown samples. The techniques proposed in this thesis are designed in such a way as to enable the recognition and learning of novel location patterns.

Since protein localisation techniques are amenable to high-throughput processing, the number of proteins with known locations is increasing. The localisation data are entered into databases such as PSORTdb [178] or Organelle DB [263]. Due to the difficulty to describe subcellular locations verbally, images are recorded as well.

3.6 Summary

There exists a great variety of methods enabling the analysis of proteins and their functions. Most of them are performed *in vitro*, that is, outside their natural environment. By means of such methods, for example, the nucleotide sequence of proteins and characteristics of their structure can be determined, but dynamic processes occurring in living organisms are not observable. However, this knowledge can be exploited by a collection of computational tools available. Unfortunately, their accuracy is rather limited. But they are capable of yielding valuable hints for practical experiments. These problems are partly solved by the yeast two-hybrid system. Nevertheless, it allows only for interactions in a cell's nucleus to be investigated. In contrast, the subcellular localisation of fluorescently-labelled proteins in intact cells enables the observation of proteins in their natural habitat. So, even dynamic processes are observable.

Owing to the extremely high number of existing proteins, automated methods are required. By means of such methods high-throughput processing becomes possible. Unfortunately, most experimental methods are not amenable to high-throughput processing of complete proteomes to date; but there exist techniques that are. From a computational point of view, here, especially an

automated evaluation of experimental data is necessary. With respect to the promising field of location proteomics, there is a need for efficient and accurate image analysis methods. Within the scope of this thesis, such techniques have been developed. They are introduced in Chapter 5 and Chapter 6.

4 Microscopy Techniques

The goal of analysing the subcellular locations of proteins in living cells necessitates special microscopy techniques. Therefore, this chapter gives a comprehensive overview of common techniques. Firstly, the properties of light, which are important within the context of microscopy, are reviewed (see Section 4.1); basic features of electromagnetic radiation have already been discussed in Section 3.1. Then, common light microscopy methods are addressed in Section 4.2. They constitute the basis for the cell recognition and the protein localisation approaches introduced in my thesis. Afterwards, more advanced techniques concerned with three-dimensional images are discussed in Section 4.3. Finally, electron microscopy, as a popular method for the detailed study of cell structures, is addressed in Section 4.4.

4.1 Properties of Light

Common light, such as sun-light, comprises a variety of photons with different energies. Therefore, its composition is important. Here, light is referred to as *monochromatic* if all waves have the same wavelength, and *polychromatic* otherwise [144, Chapter 2]. Besides this, some other properties are crucial with respect to microscopy.

The *polarisation* describes the relation of the vibration orientations of the considered photons. Linearly polarised light consists of waves whose electric fields vibrate in planes parallel to each other. By means of a superimposition of two waves, alternative kinds of polarised light can be generated: circularly polarised light and elliptically polarised light [144, Chapter 8]. These waves rotate around their propagation axis instead of vibrating in a plane. A projection of these waves on a plane perpendicular to their direction of propagation leads to circular or elliptical shapes.

A light wave's brightness can be described by the amplitude of its electric field's vibration. The phase corresponds to the retardation or expedition of two waves with respect to each other. In the case that polarised waves have the same wavelength and phase relationship, they are *coherent*.

Light is able to interact with matter in numerous ways [158]. Firstly, it might be *absorbed*. Here, the photons' energies are transformed into other kinds of energy, such as heat. The fraction of the absorbed energy is denoted by the absorption coefficient. Secondly, light can be diffracted. The *diffraction* is a deviation from the straightforward movement caused by obstacles. It is a wave-specific property that is well-known from sound waves and water waves. The diffraction caused by irregular small particles is often referred to as scattering. *Refraction* constitutes the third way of interaction. Here, the direction of a light beam is modified because of the transition from one transparent material to another. It depends on the refractive indices of the involved materials, which are tantamount to the relation of the speed of light in vacuum c and the speed of light in the respective medium. Finally, incident light may be *reflected*, that is, it cannot enter a certain material and is diverted back into the original medium.

4.2 Common Light Microscopy Techniques

Light microscopy has been of the utmost importance for the investigation of microbiological objects and microbiological research in general [133, Chapter 4]. It is still widely applied as a basic tool. Therefore, this section introduces the most essential techniques. Since these methods utilise a broad beam of illuminating light, which is focused by a condenser lens on the specimen, they are called *wide-field techniques* [167, Chapter 5].

4.2.1 Bright-Field Microscopy

Bright-field microscopy is a simple method that is frequently applied in biology [133, Chapter 4]. Due to absorption and scattering, which decrease the amplitude of incident light, a visible image of specimens is created (see Figure 4.1). Living cells are usually not harmed. The magnification is realised by two successive lenses: the *objective* and the *ocular*. In case the image is recorded, only the magnification of the objective is relevant. An additional lens, the *condenser*, is applied to focus the light from a light source on the specimen.

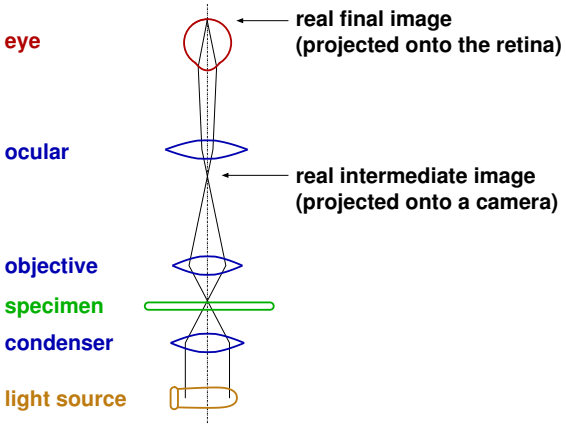


Figure 4.1: *Light path through a bright-field microscope.* The condenser focuses the light generated by a light source on the examined objects which absorb or scatter it. So, an image is generated that is projected onto the retina or recorded by a camera.

Bright-field images are intensity-variant and the contrast between observed objects and their medium is relatively low [266]. Pigmented organisms constitute an exception, as their colour adds contrast. Unfortunately, the considered cell lines, Sf9 and S2R+ (see Chapter 5 for an introduction), are not pigmented. In order to compensate for this disadvantage, dyes could be applied to stain the cells [133, Chapter 4]. But, if dyes were utilised with living cells, they might interfere with examined proteins or even kill the cells. Consequently, the application of dyes is not always reasonable. In addition, bright-field images may show a great variety of cell appearances [131]. Therefore, an automatic analysis of these images constitutes a difficult task.

4.2.2 Dark-Field Microscopy

Dark-field microscopy is a technique which is especially useful for the observation of very small specimens, such as bacteria and isolated organelles [144, Chapter 7]. In contrast to bright-field microscopy, only scattered rays which have interacted with an object are visualised; nondiffracted rays that have not been deviated by a specimen are ignored. As the image background does not contain anything that could diffract light, it is completely dark, whereas objects appear bright.

In order to obtain dark-field conditions, direct light must be prevented from being collected by the objective lens. This can be achieved by applying a dark-field condenser annulus (see Fig-

ure 4.2). An alternative method for obtaining dark-field images consists in the usage of special dark-field condensers which might be incompatible with other microscopy techniques.

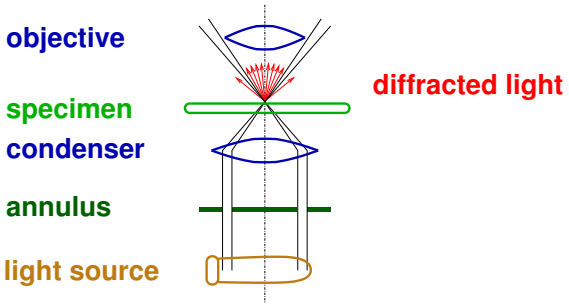


Figure 4.2: *Light path in a dark-field microscope.* An opaque plate with a transparent ring, the dark-field condenser annulus, is applied to prevent direct light from reaching the objective. As a result, the final image is composed of diffracted rays only.

Since the contrast is very high, the observed objects should be very simple [167, Chapter 5]. Otherwise, overlapping objects would scatter too much light and the image could not be interpreted. On the other hand, structural details, which might be even smaller than the resolution limit of the microscope, can be visualised [144, Chapter 8].

4.2.3 Phase Contrast Microscopy

Besides dark-field microscopy, *phase contrast microscopy* can be employed to obtain images of high contrast. As opposed to bright-field microscopy, which is based on amplitude differences of light passing through a specimen, phase contrast microscopy visualises the phase shift induced by the interaction with objects varying in thickness or refractive index [167, Chapter 5]. So, high-contrast images of living cells [133, Chapter 4][182, Chapter 1], minute organisms and even cell organelles [144, Chapter 7] can be obtained. Since the human eye is not capable of recognising phase shifts directly, they have to be transformed into amplitude differences.

In order to transform phase differences into amplitude differences, a special optical design is necessary which enables destructive interference[1] between direct and diffracted rays (see Figure 4.3). This optical design was invented by the Dutch physicist Frits Zernike [276] whose contributions are relevant for further parts of this thesis, as well (cf. Section 6.2.1).

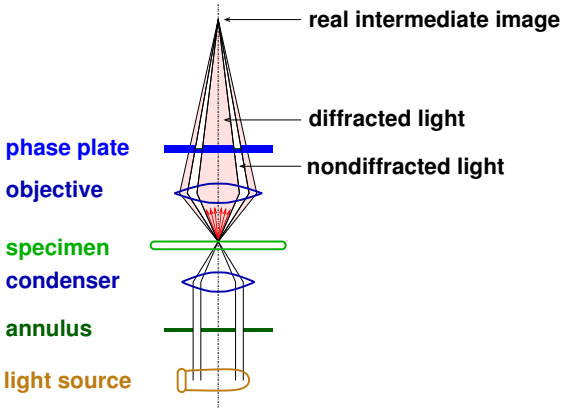


Figure 4.3: *Light path of a phase contrast microscope.* A condenser annulus is employed in conjunction with a phase plate so as to visualise phase shifts of light interacting with the specimen in comparison to direct light.

Like in dark-field microscopy, a condenser annulus is applied which allows for the separation of diffracted and undeviated light. The diffracted rays are retarded in phase relative to the non-diffracted light [133, Chapter 4][182, Chapter 1]. However, a further phase shift might be required

[1] interference which decreases the amplitude of resulting light

so as to enable destructive interference. This is commonly performed by applying a phase plate which consists of a glass plate with a ring of reduced thickness. So, the nondiffracted light is advanced. Additionally, the amplitude of the direct light must be decreased by 70% to 75% to optimise the contrast. This is accomplished by a semitransparent metal film which is coated on the ring of the phase plate.

4.2.4 Polarisation Microscopy

Besides the phase and amplitude of light, its polarisation is applicable so as to obtain images. *Polarisation microscopy* [144, Chapter 9] is especially well-suited for the visualisation of ordered molecular structures, since polarised light is able to interact with bonds of these molecules depending on their direction. Such ordered molecules exist in several biological structures, for instance in plant cell walls, chromosomes and chloroplasts. This enables further investigations, for example the analysis of the effect of drugs on living cells.

In comparison to a common bright-field microscope, a polarisation microscope comprises at least two additional components: a polariser and an analyser (see Figure 4.4). Both polariser and analyser transmit only rays with a specific plane of vibration. Here, the transmission axis of the analyser is rotated with respect to the polariser. So, only rays that have interacted with the specimen are utilised for the visualisation.

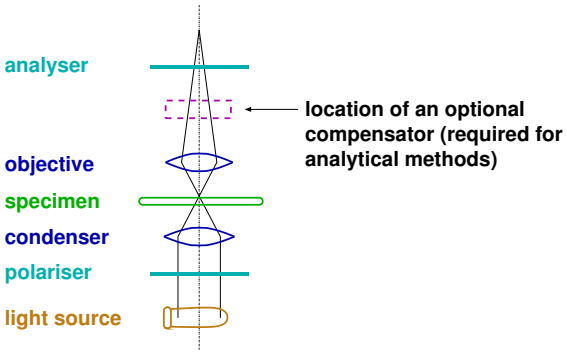


Figure 4.4: *Light path through a polarisation microscope.* The optical design resembles the bright-field microscope. But as polarised light is applied, a polariser is mounted in front of the condenser. In order to visualise interactions between the light and the specimen, a second polariser – the analyser – is necessary.

For quantitative measurements, an additional component is required – the compensator. A compensator increases the contrast significantly. So, for instance, minute retardations of rays by a specimen can be examined.

4.2.5 Differential Interference Contrast Microscopy

Differential interference contrast microscopy was developed by the Polish-born French physicist Georges Normarski [153]. Unlike phase contrast microscopy, which visualises differences between light exhibiting object-induced phase shifts and undeviated reference rays, differential interference contrast (DIC) microscopy displays local gradients of phase shifts. These gradients measure changes in the refractive index and thickness between two neighbouring points.

The determination of phase gradients necessitates a beam splitter which is realised by a prism in conjunction with polarised light. The prism is made of double refracting or rather birefringent material. So, incident polarised light is split into an ordinary ray according to the normal refractive index and an extraordinary ray, which vibrate in perpendicular planes and are separated by a small distance (0.2–2μm). After traversing the specimen, these rays are recombined by a second prism.

The visible image is then created by the analyser whose transmission axis is rotated 90° with respect to the polariser (see Figure 4.5).

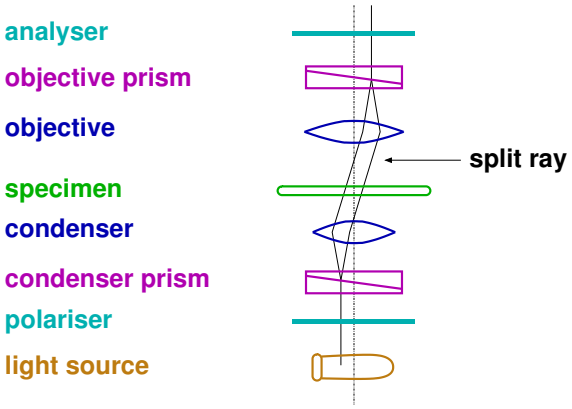


Figure 4.5: *Light path through a differential interference contrast microscope.* The condenser prism splits polarised light into two beams which traverse the specimen. The following recombination by the objective prism leads to an elliptically polarised ray being partially transmitted by the analyser. The interference of both components in the analyser results in an image depicting phase gradients.

In the case of a phase difference between an ordinary and an extraordinary ray, the recombination leads to elliptically polarised light which partially passes the analyser. As the resulting light is linearly polarised, interference occurs that visualises the phase gradient. If there is no phase difference, linearly polarised light results that is blocked by the analyser causing a black background. In order to enhance the image contrast, a bias retardation is usually introduced. So, a phase shift between ordinary and extraordinary rays is induced. As a result, the background appears bright, whereas gradients are shown dark or bright depending on the sign of the gradient.

4.2.6 Fluorescence Microscopy

Provided that a specimen contains fluorescent molecules, it can be examined by means of *fluorescence microscopy*. The principal design of a fluorescence microscope is depicted in Figure 4.6. The produced images show a very high contrast.

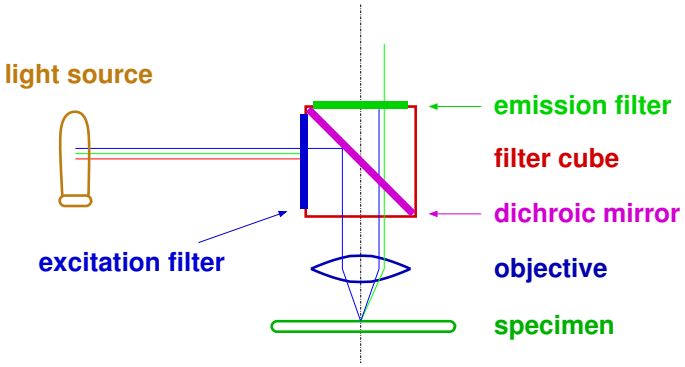


Figure 4.6: *Light path in a fluorescence microscope.* Light generated by a light source is filtered so as to block wavelengths not required for excitation. This filtered light is reflected to the specimen by a dichroic mirror. After the excitation of fluorescent molecules, the resulting light is transmitted by the same dichroic mirror. Additionally, the emission filter blocks short-wavelength light which might harm a viewer. Both filters and the dichroic mirror are arranged in a filter cube. Remarkably, the objective acts as both condenser and objective.

Unfortunately, most specimens that ought to be observed by fluorescence microscopy are not inherently fluorescent. Thus, the usage of fluorescent dyes is often inevitable. These dyes can be applied through utilising a great variety of methods [133, Chapter 4] (cf. Section 3.5). DAPI,

for example, binds to nucleic acids and enables the observation of cell nuclei. Moreover, specific organisms and intracellular structures can be visualised by employing fluorescent antibodies which are to be designed in advance. Another technique consists in the construction of visible fusion proteins with the green fluorescent protein (GFP) [235] (see Section 3.5 as well).

4.3 Three-Dimensional Imaging

Biological specimens are inherently three-dimensional objects. Therefore, their tree-dimensional structure should be considered instead of a two-dimensional projection like in wide-field microscopy. Then, not only light from focal planes below and above the current focus can be eliminated, but the complete appearance of the objects is revealed as well. Several methods have been proposed so as to enable three-dimensional imaging. Digital deconvolution (cf. Section 4.3.1), for example, constitutes a mathematical approach, which can be applied in conjunction with common wide-field microscopes. In contrast, confocal laser scanning microscopy (see Section 4.3.2) utilises a special optical design which enables the scanning of single points in a three-dimensional raster. In principle, both methods necessitate more time for the acquisition of an image than conventional wide-field microscopy. In contrast to confocal laser scanning microscopy, spinning disk microscopy (see Section 4.3.3) is significantly less time consuming. The intensities of a two-dimensional image can be scanned simultaneously. Thus, this method is especially well-suited for the analysis of live cells.

4.3.1 Digital Deconvolution

The quality of images is limited by the performance of optical microscopes. The depth of focus, in particular, is lacking in precision. Therefore, every image contains information from adjacent focal planes, which contaminate it. Because of the imperfect optical properties, a single point of the specimen is spread over a finite volume. This volume is characterised by the *point spread function* (PSF). If the PSF is known, the original image can be reconstructed by *digital deconvolution* [4][255] from a stack of images taken at adjacent focal planes.

Several approaches such as nearest-neighbour deconvolution, constrained iterative algorithms and statistical methods have been proposed in order to tackle this problem. A comprehensive overview of these techniques is given in [200]. Digital deconvolution usually requires an estimate of the PSF so as to allow for a reasonable reconstruction of the original image. It can be determined theoretically, depending on the microscope design, or measured empirically. Both methods yield suboptimal results; theoretical models describe idealised lenses, which do not exist, and measured PSFs may be disturbed by noise. Furthermore, deconvolution algorithms make additional suppositions. Frequently, linearity and shift invariance are assumed for the image creation process in order to enable the usage of convolution filters and the Fourier transform [74, chapters 3 and 4]. So, despite an additional computational effort, it cannot be guaranteed that the restored images depict real biological entities. Nevertheless, digital deconvolution is able to significantly increase the quality of the images. Therefore, it is extensively applied within the context of protein localisation [36][89][112].

4.3.2 Confocal Laser Scanning Microscopy

Besides digital deconvolution, *confocal laser scanning microscopy* [144, Chapter 12][167, Chapter 5] can be applied so as to decrease the amount of incident light from objects outside the focal plane. The foundations of this technique were laid by the American Marvin Minsky at Havard University in 1957 [143].

The removal of light from out-of-focus objects is achieved by utilising a focused laser beam and a pinhole aperture; the laser beam limits the amount of diffracted light within the illuminating beam and the pinhole aperture eliminates light from out-of-focus planes. So, the light intensity from a specific point of a specimen can be recorded. In order to obtain an image, the whole specimen has to be scanned in a raster pattern point by point. The final image is created artificially by a computer which assembles the scans of all points. The three-dimensional structure of biological objects can be visualised by performing a series of scans at different focal planes.

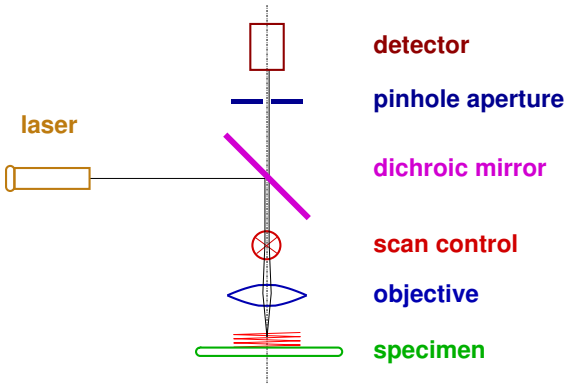


Figure 4.7: *Optical path in a confocal laser scanning microscope.* The specimen is point-wisely scanned by a focused laser beam. In contrast to other microscopy methods, the image is not directly visible. The detector outputs are rather composed by a computer.

Confocal laser scanning microscopy can be employed in order to investigate fluorescent specimens as well as objects which reflect light. Therefore, additional components such as emission and excitation filters must be added to the microscope design. Due to the intensity of the laser beam, the amount of fluorescent molecules might be decreased as a result of chemical modifications. This effect is called *photobleaching* and should be taken into account.

4.3.3 Spinning Disk Microscopy

In order to overcome the speed limitations of confocal laser scanning microscopy, the specimen can be scanned by multiple beams, which has been realised by *spinning disk microscopy* [79][144, Chapter 12]. For this, the laser must be widened. Furthermore, a spinning disk based on the Nipkow disk, which was developed by the German inventor Paul Nipkow in 1884, has to be utilised. So, the laser beam is split, which allows for the simultaneous scanning of multiple points. Hence, a single detector is not sufficient, but rather a camera has to be employed (see Figure 4.8).

Besides significantly decreasing the time required for taking an image, considerably less excitation light is necessary for the acquisition of fluorescence images [79]. So, negative effects like photodamage caused by light-induced chemical reactions and photobleaching can be reduced. Therefore, this microscopy method is particularly well-suited for the analysis of live cells.

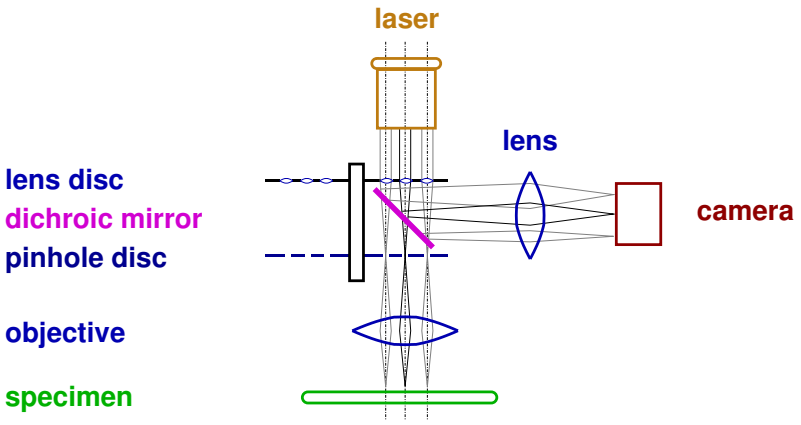


Figure 4.8: *Path of light in a spinning disk microscope.* Multiple points of a specimen are scanned in parallel by splitting the illuminating laser beam using a pinhole disk and a lens disk instead of a single objective and a single pinhole aperture (cf. Figure 4.7). The intensities of the resulting beams are measured independently by a camera.

4.4 Electron Microscopy

Common light microscopes cannot reach resolutions smaller than approximately 200nm, as they are constrained by the wavelengths of visible light [58]. Therefore, alternative microscopy techniques are required. One possibility to obtain a better resolution consists in the usage of electromagnetic waves of higher frequency, such as ultraviolet light. The application of beams of electrons, which is realised by means of *electron microscopy*, leads to similar effects. According to a theory developed by the French mathematician Louis de Broglie in 1924 [49] an electron beam can be considered as a wave. The corresponding wavelength λ results from Equation 4.1 where h denotes Planck's constant, m the mass of an electron, and v its velocity.

$$\lambda = \frac{h}{m \cdot v} \tag{4.1}$$

The resulting wavelengths are orders of magnitudes smaller than the ones of visible light. In principle, atomic resolution is possible. However, the resolution with respect to biological specimens only reaches values of approximately 2nm, since the preparation of specimens imposes additional limits [58].

Images are created as the result of two effects: elastic and inelastic scattering [58]. Elastic scattering describes the deflection of electrons by the nuclei of atoms. Here, larger nuclei cause a greater deflection. If the electron beam interacts with an orbital electron instead of the nucleus of a specimen's atom, inelastic scattering takes place. Here, the electrons' energy is reduced. After passing a specimen, the electron beam is projected onto a fluorescent screen which emits photons depending on the energies of incident electrons.

Due to the significantly improved resolution, special requirements with respect to the preparation of specimens have to be fulfilled [52]. Firstly, the specimen must be fixed. Secondly, the application of a vacuum necessitates the considered objects to be dehydrated or frozen. Finally, stains must be utilised frequently, since biological structures exhibit low contrast. These stains usually consist of heavy metals.

5 Cell Recognition

In order to localise proteins in cell compartments, fluorescence microscopy is employed (cf. Section 3.5): The proteins are tagged by a fluorescence dye, e.g. the green fluorescent protein. So, cell compartments containing these proteins are visualised as bright image regions (see Figure 5.1).

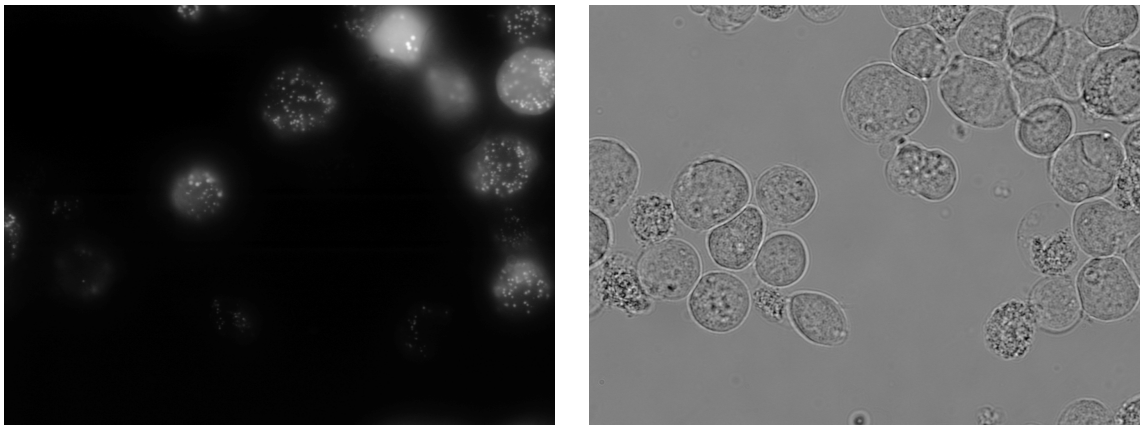


Figure 5.1: *Fluorescently labelled proteins which are localised in the peroxisomes of Spodoptera frugiperda cells (a) and a corresponding bright-field image (b).* In comparison with the bright-field image, it becomes clear that the cells cannot be recognised solely based on the depicted fluorescence micrograph.

Unfortunately, the cells are not necessarily visible in fluorescence micrographs. Hence, the positions of the surrounding cells must be determined first. But the subcellular localisation of proteins imposes special limitations to any potential cell recognition method. Firstly, such an approach must not influence the result of an investigation. Secondly, the quality of the applied fluorescence micrographs should not be decreased. Thus, only a small number of the published cell recognition methods is applicable in conjunction with protein localisation methods. Before these cell recognition techniques are reviewed (see Section 5.2), the employed cell line (Sf9), which originates from the moth *Spodoptera frugiperda*, is introduced in Section 5.1.

The subsequent sections describe the proposed cell recognition system. In order to find Sf9 cells in microscope images in an automated way, several tasks have to be fulfilled (cf. Figure 5.2). At first, possible cells must be localised, i.e. the positions of such candidate cells are to be determined. Here, several intermediate images are computed: one image showing possible cell membranes and another one depicting the image background. Based on these two images, small regions within the possible cells are determined – the *cell markers*. They reflect the positions of the surrounding cells. Unfortunately, at this step no differentiation between real cells and other image objects is possible, since too little information about the corresponding image objects is available.

After the localisation of candidate cells, they are segmented – that is, all pixels showing a specific cell are associated with it. Then, representative features describing a cell are computed and non-cell objects can be rejected. This is achieved by means of a classifier.

The localisation and segmentation of possible cells are discussed in Section 5.3. Afterwards,

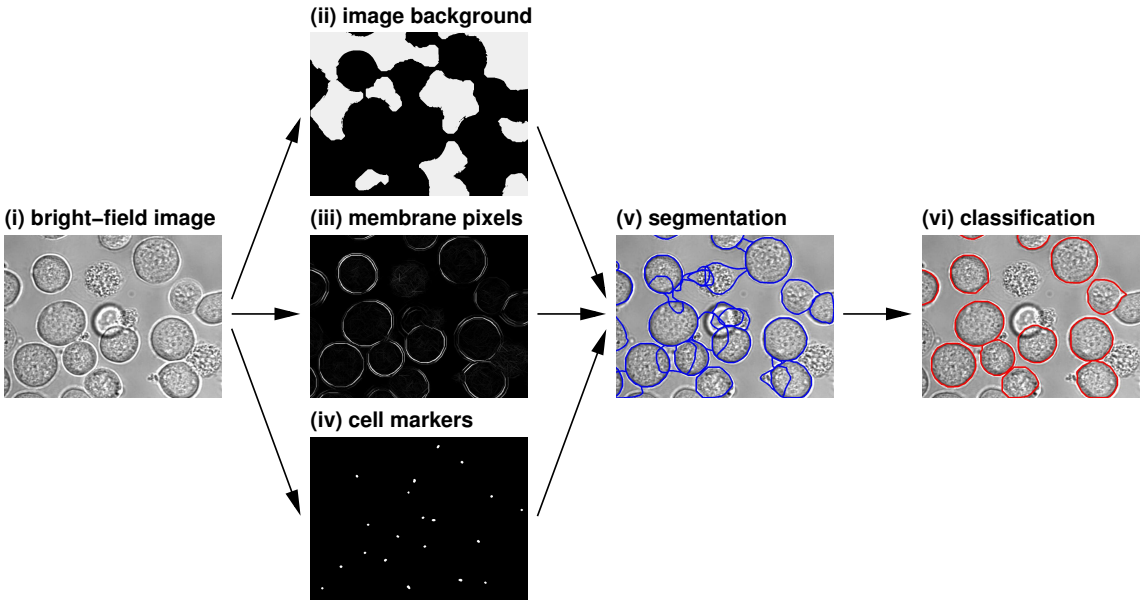


Figure 5.2: *Outline of the proposed cell recognition approach.* On the basis of an acquired bright-field image (i) three further images which depict background pixels (ii), probable cell membrane pixels (iii), and cell markers (iv) are generated. They constitute the foundation of the proposed segmentation procedure. The segmentation (v) is followed by a classification step (vi) rejecting non-cell segments.

Section 5.4 addresses the rejection of non-cell segments. The performance of the complete system is evaluated in Section 5.5. Afterwards, an adaptation of the proposed approach to an alternative cell type (S2R+) originating from *Drosophila melanogaster* is presented (see Section 5.6). Finally, the most important results are briefly summarized in Section 5.7.

I have already published parts of the proposed approaches. Details regarding the localisation and segmentation of Sf9 cells appeared in [231], whereas their classification is addressed in [233]. One of the used classification techniques, which constitutes a modification of the simplified fuzzy ARTMAP [241], is detailed in [228]. The first results using an alternative magnification were discussed in [232]. Additionally, a comparison of several segmentation techniques is given in [234]. In contrast to the Sf9-specific cell recognition method, the technique enabling a recognition of S2R+ cells was completely published in a single article [228]. But although the principal techniques have already been introduced, the majority of the evaluations was performed anew in order to reach a better continuity of this thesis.

5.1 The Employed Cell Line

Although prokaryotes are able to express foreign proteins (cf. Section 3.4.2), they cannot perform numerous post-translational modifications that are necessary for the correct function of proteins from higher organisms, especially mammals and humans. Therefore, alternative systems are required in order to analyse such proteins, which constitute the basis for the investigation of novel diagnostic or therapeutic targets. As a result, the economical interest in insect and mammalian cells has increased [75][223].

Insect cells have been proven to be beneficial for the high-level expression of foreign proteins [75][85][114][123]. Here, the proteins are often correctly modified and localised. Genetic engineering enables the proper processing of additional proteins [224]. Besides their application to pro-

tein expression, insect cells have been studied with respect to insect pest management [123]. As a result, a large number of cell lines originating from several insect species are available [75][245].

The cells, the microscope images of which are analysed within the scope of this thesis, stem from the fall army worm *Spodoptera frugiperda* – a moth inhabiting the northern hemisphere. In 1977 a cell line called IPLB-SF-21 was extracted from immature ovaries of *Spodoptera frugiperda* pupae [245]. It served as a basis for the derivation of a further cell line termed *Sf9* [211].

Sf9 cells have several beneficial features, which make them amenable to high-throughput investigations [105][183]. First of all, they are robust and not very demanding; for example, they grow at room temperature and in serum-free medium without any added growth factors [54]. In addition, Sf9 cells exhibit a round shape with diameters between about 15μm and 20μm which is relatively large. Cells from the budding yeast *Saccharomyces cerevisiae*, for example, reach only diameters of about 8μm [133, Chapter 4]. So the differentiation between protein localisation patterns is alleviated if Sf9 cells are employed. In contrast to mammalian cells, the cell growth is independent from carbon dioxide. As a result, no special devices are required. Finally, they are adherent, i.e. they form a single layer attached to a surface. Through this, the application of automatic techniques such as auto-focus or image analysis procedures is facilitated.

5.2 Related Work

In order to account for the limitations, which have to be considered within the context of automatic protein localisation in living cells, Section 5.2.1 introduces and evaluates basic microscopy techniques frequently used in conjunction with cell recognition approaches. As the choice for a recognition method strongly depends on the utilised microscopy technique, the application of several well-known approaches, which are discussed in Section 5.2.2, is partly impeded.

5.2.1 Evaluation of Microscopy Techniques

A large number of cell recognition approaches such as [51][175][278] employ phase contrast microscopy (see Section 4.2.3) to increase the contrast of acquired images. It visualises the phase shift induced by the interaction with objects varying in thickness or refractive index. Since this microscopy technique requires special objectives that reduce the amplitude of incident light, the light from fluorescent objects would be attenuated as well. An alternation of the objective between the acquisition of the images used for protein localisation and cell recognition causes further problems, since it modifies the optical path. Consequently, an association of corresponding pixels of these images would be hampered.

Besides phase contrast microscopy, numerous approaches require special dyes (see Section 4.2.6) [127][151][173][177][190]. If they were used within the scope of a protein localisation approach, they might interfere with the examined proteins or influence the cell state.

Bright-field microscopy (cf. Section 4.2.1), i.e. the direct observation of illuminated objects, is a widely used method for cell observation. It is usually available without any special devices. But the resulting contrast is rather low, which necessitates more complex recognition techniques [130][131][231][233][266].

Differential interference contrast (DIC) microscopy (see Section 4.2.5) displays local gradients of the phase shift between two neighbouring points. The resulting images may exhibit a better contrast in comparison to bright-field images, but are more difficult to interpret [271][272]. The

optical gradient results in transitions from bright to dark pixels giving objects an almost three-dimensional appearance, which does not necessarily correspond to their real shape.

Dark-field microscopy (cf. Section 4.2.2) is a technique which is especially useful for the observation of very small specimens, such as bacteria and isolated organelles [144, Chapter 7]. In contrast to bright-field microscopy, only scattered rays, which have interacted with an object, are visualised; nondiffracted rays that have not been deviated by a specimen are ignored. In order to obtain dark-field conditions, direct light must be prevented from being collected by the objective lens, which might result in incompatibility with other microscopy techniques. Nevertheless, it enables additional conclusions about the cells such as their viability [257].

On the whole, bright-field microscopy is probably the most frequently applied microscopy technique. Therefore, I have decided to use bright-field images as the basis of the cell recognition method. In addition, differential interference contrast (DIC) images have been employed, since they might reveal details that are not visible in bright-field images. Both techniques can be used in conjunction with fluorescence microscopy. Similar combinations of images obtained by means of different microscopy techniques have already been applied successfully [76].

5.2.2 Well-Known Cell Recognition Approaches

The most common approach to cell recognition probably consists in thresholding [37][76]. But it is often applied to cell nuclei rather than whole cells [134][267][252]. As each cell usually has a single nucleus, which covers a large fraction of its volume, these tasks are roughly equivalent.

Thresholding requires a uniform and unambiguous distribution of pixel intensities, which does not occur either in bright-field or in DIC images that show a great variety of cell appearances. Even if fluorescence images of stained nuclei are to be analysed, fuzzy transitions between objects and the image background may result in difficulties in selecting a proper threshold. In addition, thresholding causes problems in separating adjoining objects, which have to be dealt with separately. Here for example, the distance transform and the *watershed transform* can be applied [134][267] (cf. Section 5.3.5). Nevertheless, the prior binarisation of the image leads to a loss of information, which might be crucial for the determination of the objects' exact boundaries.

As an alternative to thresholding, there are approaches that determine and link the edges of stained nuclei using geometrical constraints [173]. Unfortunately, these constraints do not necessarily reflect the shape of visible objects – especially if these objects partially overlap.

Since subcellular structures are to be analysed after cell recognition, a high magnification is required. Hence, I decided to apply images taken at $60\times$ magnification, which is the maximum magnification of the employed microscope. So, the considered cells comprise between 10,000 and 80,000 (Sf9), and 3,000 and 25,000 pixels (S2R+), respectively. As a result, methods utilising small rectangular patches in order to detect whole cells (cf. [130][131][151][190]) cannot be employed, as the computational costs would be too high. So, for example, the approach proposed in [130] takes 1 to 8 minutes in order to recognise cells in relatively small images (640×480) using a patch size of 625 pixels on an Intel Pentium 4 processor operating at 1.6GHz.

However, Petra Perner and her co-researchers proposed a technique for recognising fungal spores using bright-field microscopy [162], which resorts to image pyramids in order to decrease the processing time. The suggested technique iteratively compares small image regions with a set of examples for the objects under consideration, referred to as cases. Since these cases constitute images themselves, the translation, rotation and scaling of the cells have to be dealt with

explicitly. A more abstract representation, for instance by means of representative features, could circumvent these problems. Furthermore, it might allow for a better generalisation, since irrelevant information can be neglected. Nevertheless, the given recognition rates appear very promising.

Cells in bright-field and DIC microscope images are separated from other cells and the surroundings by their membrane. Consequently, it is beneficial to include information about it in the segmentation procedure. This can be accomplished by determining cell membrane pixels and linking them [6][272]. But, in the case of images containing numerous cells of varying shape or size, it is difficult to obtain unambiguous solutions.

As an alternative to edge-linking methods, *snakes* (cf. Section 5.3.5) have proven advantageous [175][231][278]. Besides exploiting gradient and image information, they allow for the incorporation of prior knowledge on cell features such as curvature and size without assuming a rigid model. Therefore, I decided to develop a snake-based algorithm for the recognition of Sf9 cells in bright-field images [231][232][233][234]. This algorithm has been extended to enable a recognition of an alternative cell type called S2R+ as well [228]. But here, pairs of corresponding bright-field and DIC images are utilised.

Besides snakes, level set methods have been applied to cell recognition tasks [50]. Due to their close relationship with snakes, they will be discussed in Section 5.3.5 in more detail.

5.3 Localisation and Segmentation of Probable Cells

The first step of the proposed cell recognition system consists in the localisation of possible cells. In order to fulfil this task, morphological features of the observable cells are exploited. Hence, morphological operators were employed, as they enable the investigation of image structures with respect to specific shapes. The operators relevant for this thesis are introduced in Section 5.3.1. As morphological operators are critical for the understanding of several parts of my thesis, they are explained in detail. The subsequent sections 5.3.2, 5.3.3 and 5.3.4 are devoted to the actual localisation approach. Here, the image background is separated from the image foreground, possible cell membrane pixels are selected and cell markers are determined, respectively.

Based on the cell markers, the images are segmented. Therefore, relevant segmentation procedures are discussed in Section 5.3.5. From these methods, parametric active contours appeared most promising, because of which they were chosen. Section 5.3.6 introduces the basic properties of this method. Furthermore, two different techniques for its practical realisation are considered. Finally, the proposed techniques are evaluated by means of a set of manually segmented cells (see Section 5.3.7).

5.3.1 Morphological Operators for Grey-Scale Images

Mathematical morphology is a powerful tool for performing image analyses [206]. It enables the investigation of images structures concerning their shapes. Hence, if an image contains objects of a specific shape which are to be analysed, the application of morphological operators is beneficial.

The first works on mathematical morphology were constrained to binary images [139, Chapter 1]; a binary structuring element was applied so as to probe the image structure. Based on these approaches, several extensions to grey-scale images were proposed. The most important techniques are the threshold approach, the umbra approach and methods derived from fuzzy set theory

[149]. Threshold procedures apply binary or rather *flat structuring elements* resulting in independence of the shape of the structuring elements from the scaling of the grey values. In contrast, the latter two methods utilise grey-scale or *nonflat structuring elements*, which depend on the intensity scale of an image. In order to take advantage of the independence from the scaling of image grey values, I decided to employ the threshold approach.

The basic operators provided by mathematical morphology are the *erosion* and the *dilation*. Their theoretical background was already examined at the beginning of the 19th century, e.g. by Hermann Minkowski who investigated mixed volumes of three-dimensional solids [142]. The erosion $\varepsilon_S(I)$ by a structuring element S is defined according to Equation 5.1. Here, $\underline{\mathrm{x}}$ denotes a point from the image I and $\underline{\mathrm{s}}$ a point of the structuring element S. The origin of S must be placed at $\underline{x}$.

$$\big[\varepsilon_S(I)\big](\underline{x}) = \min_{\underline{s}\in S} I(\underline{x}+\underline{s}) \tag{5.1}$$

The dilation $\delta_S(I)$ is defined accordingly using a maximum rather than a minimum operation (see Equation 5.2).

$$\big[\delta_S(I)\big](\underline{x}) = \max_{\underline{s}\in S} I(\underline{x}+\underline{s}) \tag{5.2}$$

More powerful morphological operators can be constructed by combining erosions and dilations. So, the basic set of operators can be extended by the *opening* and the *closing*. An opening $\gamma_S(I)$ removes bright image structures which cannot include the structuring element S. It is computed by consecutively performing an erosion with S and a dilation with the reflected structuring element $\check{S}$ (see Equation 5.3).

$$\gamma_S(I) = \delta_{\check{S}}\big[\varepsilon_S(I)\big] \tag{5.3}$$

In contrast to an opening, a closing eliminates dark image structures which are not able to include S. It is defined according to Equation 5.4.

$$\phi_S(I) = \varepsilon_{\check{S}}\big[\delta_S(I)\big] \tag{5.4}$$

A further extension of the operator set is achieved by geodesic transformations which are applied to two input images. The second image M serves as a mask that defines either an upper or a lower bound of the result. In addition, no specific structuring element must be chosen, since only erosions and dilations with an elementary structuring element σ comprising a single pixel and its direct neighbours are allowed.

The *geodesic dilation* of size 1 is defined by Equation 5.5. Equation 5.6 shows the definition for an arbitrary number of iterations n. Here, $\wedge$ symbolises the computation of a point-wise minimum.

$$\delta_M^{(1)}(I) = \delta_\sigma(I) \wedge M \tag{5.5}$$

$$\delta_M^{(n)}(I) = \delta_M^{(1)}(I)\big[\delta_M^{(n-1)}(I)\big] \tag{5.6}$$

The *geodesic erosion* is defined similarily (see Equations 5.7 and 5.8). Here, the point-wise maximum operator $\vee$ is employed so as to incorporate the mask image M.

$$\varepsilon_M^{(1)}(I) = \varepsilon_\sigma(I) \vee M \tag{5.7}$$
$$\varepsilon_M^{(n)}(I) = \varepsilon_M^{(1)}(I)\left[\varepsilon_M^{(n-1)}(I)\right] \tag{5.8}$$

One of the most powerful morphological operators – the *morphological reconstruction* – can be obtained by iterating this transformation until stability. Depending on the chosen geodesic operation, the morphological reconstruction is defined as follows:

$$R_M^\delta(I) = \delta_M^{(s)}(I) \tag{5.9}$$
$$R_M^\varepsilon(I) = \varepsilon_M^{(s)}(I) \tag{5.10}$$

The morphological reconstruction constitutes the basis for numerous operators which have been applied during the work for this thesis, for instance, the computation of regional maxima, the imposition of minima and the removal of structures connected to the image border (see Section 5.3.4 and Section 5.3.5). Therefore, its computational efficiency becomes crucial. Because of that, a very fast reconstruction algorithm developed by Luc Vincent was implemented [248]. This technique is an order of magnitude faster than the original iterative approach.

5.3.2 Separation of the Image Foreground and Background

Kenong Wu and his colleagues have shown that the local intensity variation is a valuable feature for the separation of the foreground and the background in bright-field images [266]. Instead of computing the local variation defined by the variance within a square neighbourhood, I take advantage of a morphological operator: the *self-complementary top-hat* $\varrho_S(I)$ (see Equation 5.11) [206][231]. This operator preserves bright as well as dark image structures that cannot include the structuring element S.

$$\varrho_S(I) = \phi_S(I) - \gamma_S(I) \tag{5.11}$$

Figure 5.3 depicts the result of the application of the self-complementary top-hat to an exemplary bright-field image as well as the corresponding variance map if a square neighbourhood of 41×41 pixels (suggested by Kenong Wu and his co-researchers) is considered.

The bimodal distribution of the local intensity variations resulting from the application of the self-complementary top-hat is considerably more distinctive than the one computed by analysing the variance. Hence, the automatic separation of image foreground and background is alleviated. Here, minimum error thresholding [119] is utilised, as it yields excellent results for the emerging grey-level distributions [266].

In order to increase the computational efficiency, structuring elements comprising 25×25 pixels were employed. Despite their reduced size, they still performed better than the variance map using a neighbourhood of 41×41 pixels.

In principle, the application of structuring elements that do not have a rectangular shape would be possible as well. But, since rectangular structuring elements can be decomposed into two linear elements, they are more computationally efficient [206]. Furthermore, other shapes did not yield considerably improved results with respect to the task at hand.

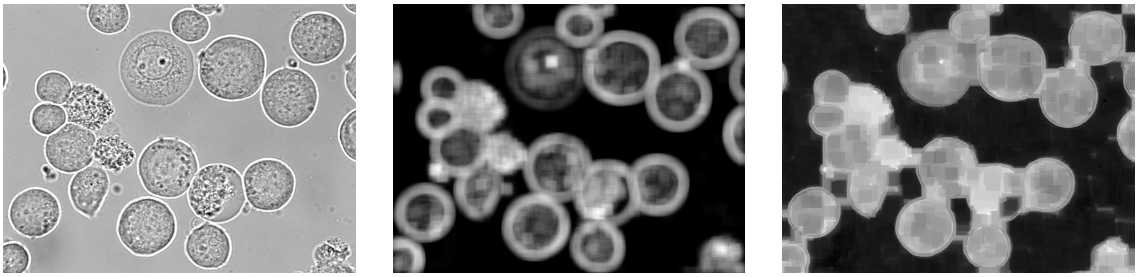


Figure 5.3: *Local intensity variations in a bright-field image (a).* The result of the self-complementary top-hat (c) allows a noticeably better recognition of the image foreground than the variance map (b) using a neighbourhood of 41×41 pixels. In particular, the variance map only enables a detection of foreground pixels close to the cell membrane, whereas the self-complementary top-hat allows for the recognition of the complete image foreground.

Even if the self-complementary top-hat is applied, the separation of the image foreground and background very close to the boundaries of cells is often incorrect, as here scattered light causes a higher variance of the local intensity values. Thus, background pixels near a cell's boundary might be associated with the image background. In order to solve this problem, a geodesic erosion is applied in such a way that the background cannot be extended in image regions exhibiting a gradient magnitude which is higher than a threshold τ_{bg}. As cell membranes cause a high gradient magnitude, the image background is prevented from growing into cells. In addition, the number of iterations n_{bg} is limited. Appendix A.1 proposes a simple method for the automatic determination of both τ_{bg} and n_{bg}. Moreover, an algorithm enabling the fast computation of the geodesic erosion is introduced there.

5.3.3 Detection of Pixels Probably Showing Cell Membranes

Probable cell membrane pixels are determined by utilising morphological operators, as well, since they enable the inclusion of knowledge concerning the shape of the image structure in question. As the cell membrane possesses a linear shape that is less curved than other cell compartments, linear structuring elements are applied. The membrane is further characterised by a substantial change of intensities between neighbouring pixels. Therefore, the gradient magnitude image is utilised instead of the original image. All image structures that cannot contain the linear structuring element, e.g. dirt, noise and intracellular objects, are removed by a morphological opening. In order to get closed contours, this operation is repeated for seven additional orientations. The resulting images are fused by computing the point-wise maximum. The whole operation constitutes an algebraic opening [206] (see Figure 5.4).

The length l of the linear structuring elements is crucial to the result of the algebraic opening. If it is chosen too small, irrelevant image structures will remain; if the value is too high, cell membrane pixels will disappear. Hence, a procedure for the automatic determination of an optimal value was developed (see Appendix A.2) [231].

In order to decrease the computational effort, an optimised method enabling the computation of morphological openings using line elements at arbitrary angles was implemented. It is based on methods proposed by Pierre Soille and his colleagues [207] who generalised an algorithm originally introduced by Marcel van Herk, which solely allowed for the usage of horizontal, vertical and diagonal lines [242]. Soille's algorithm performs morphological operations independently of

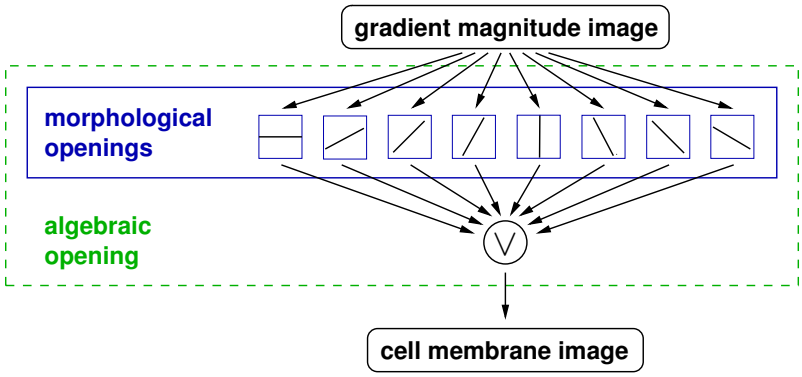


Figure 5.4: *Detection of pixels probably representing cell membranes.* Morphological openings with linear structuring elements having eight different orientations are performed to suppress image structures that do not represent cell membranes. The resulting images are fused by a point-wise maximum operation denoted by '$\vee$'.

the length of the linear structuring elements used. In particular, images of a specific size can be processed in constant time with respect to the structuring elements' lengths.

5.3.4 Determination of Cell Markers

On the basis of the computed image background and cell membrane pixels, small regions within probable cells are determined – the cell markers (see Figure 5.5). Here, only membrane pixels, the intensity of which is higher than a threshold τ_{m}, are considered. τ_{m} is set to one forth of the maximal pixel intensity so as to suppress background noise. By virtue of the assumption that points possessing a great distance to the image background and to membrane pixels lie inside cells, the distance transform [18] is applied. The local maxima of the resulting image constitute the required cell markers and can be easily computed by utilising operators based on the morphological reconstruction [206].

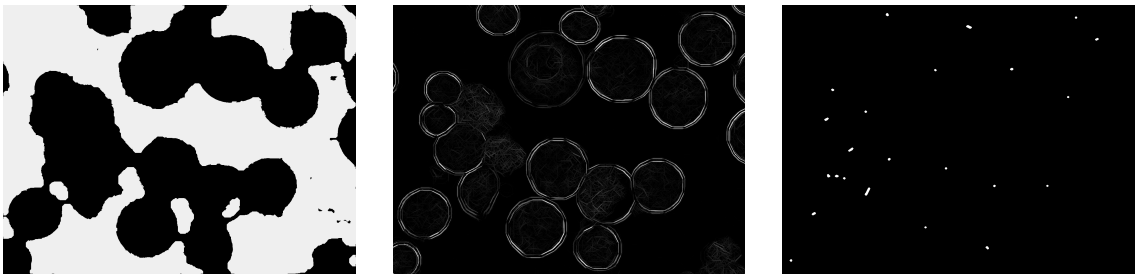


Figure 5.5: *Computation of cell markers.* The cell markers (c) are determined in such a way that they maximise the distance to the image background (a) and membrane pixels (b).

In order to obtain an appropriate initialisation for the segmentation step, these regions are dilated by a small circular structuring element with diameter d_{m}, which is set to 5% of the maximal cell radius (9 pixels). Afterwards, the contours are traced by a contour-following algorithm [103, Chapter 9]. So, a polygonal representation is determined that comprises only the start and end points of adjoining lines. The dilation is required to eliminate maxima comprising less than three points as they do not yield proper contours.

5.3.5 Comparison of Important Segmentation Methods

In order to segment Sf9 cells in bright-field images several segmentation techniques were applied. Their results differ significantly from each other – in terms of the segmentation error and in terms of the segments' characteristics.

Firstly, *seeded region growing* (SRG) was employed [2]. SRG is a region-based segmentation procedure starting from small regions representing the final segments – the seeds. These seeds, which can encompass only one pixel each, are extended according to a homogeneity criterion. At first, all pixels which do not belong to a segment but have a neighbour that does are put into a *priority queue* [204, Chapter 2], which is also referred to as *sequentially sorted list* (SSL). The priority is represented by the homogeneity criterion. After this initialisation, the priority queue is processed. Pixels best fitting the existing categories are selected first. They are associated with the adjoining segment and their neighbours are put into the queue. So, the whole image is processed.

The segmentation of a cell image is depicted in Figure 5.6. The cell markers as well as the image background were used as seeds (cf. Figure 5.3). Here, the actual seed regions and labels were determined using a labelling technique [206, Chapter 2]. So only pixels of the same connected region share a label.

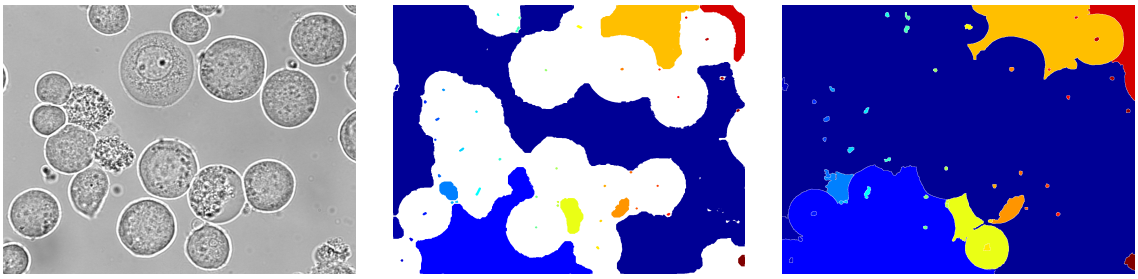


Figure 5.6: *Segmentation of a cell image using SRG.* A bright-field image showing Sf9 cells (a) was segmented by means of seeded region growing. The cell markers and the image background were used as seeds (b). Coloured regions correspond to the seeds of different regions, whereas white pixels do not belong to any seed. In the final picture (c), only border pixels between to adjoining segments are depicted white. All other pixels were assigned to a segment

Figure 5.6 clearly shows that homogeneity of the considered regions is not a suitable criterion for the segmentation of Sf9 cells in bright-field images. Although some cell borders were found correctly, no cell was properly segmented. This led to the conclusion that cell membranes are a crucial feature for the segmentation. Even more-advanced SRG approaches that apply constraints to the border of segments cannot compensate for this problem [13]. Therefore, the *watershed transform*, a hybrid technique considering region and edge information, was examined [181][206, Chapter 9].

In order to compute the watershed transform of a grey-scale image, the image can be considered as topographic representation of a landscape. This landscape is progressively flooded with imaginary water beginning with the deepest valleys (see Figure 5.7). The valleys are referred to as *catchment basins*. If water from different catchment basins merges, a dam is built. These dams are called watershed lines. They separate adjoining segments. If the complete landscape is filled with water, the procedure is terminated.

In cases where the watershed transform is applied to an image without prior processing, small catchment basins resulting from noise lead to over-segmentation. Hence, a selection of relevant

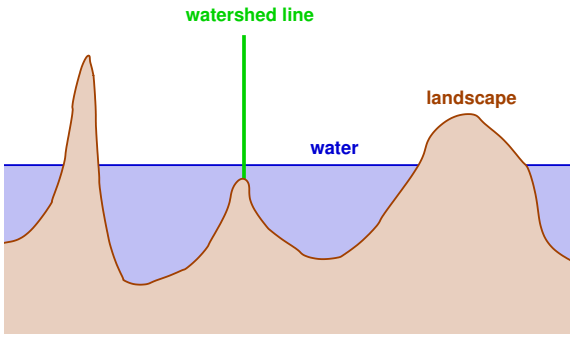


Figure 5.7: *Image segmentation using the watershed transform.* An grey-scale image regarded as three-dimensional landscape is flooded with imaginary water. The segments correspond to the landscape's valleys. They are separated by means of the watershed lines which constitute dams preventing water from different valleys from merging.

basins is beneficial. This can be achieved by means of markers similar to the seeds of SRG. Based on these markers, an image I^{m} is constructed, which equals 0 for all marker pixels and the maximum intensity otherwise. Then, I^{m} is fused with the considered image I using a morphological operator called *minima imposition* (see Equation 5.12).

$$\iota(I, I^{\mathrm{m}}) = R^{\epsilon}_{(I+1)\wedge I^{\mathrm{m}})}\big(I^{\mathrm{m}}\big) \tag{5.12}$$

At the beginning of a minima imposition $\iota(I, I^{\mathrm{m}})$, I is incremented by one. Subsequently, the point-wise minimum with the marker image I^{m} is determined. As a result, no catchment basins exist which are deeper than the ones corresponding to the markers. Finally, all spurious basins are filled using the morphological reconstruction (cf. Section 5.3.1).

The watershed transform is usually applied to the gradient magnitude image instead of the original image. Then, edges correspond to mountains of the topographic representation, whereas regions with small intensity variations form valleys. This is especially advantageous if cells in bright-field images are considered. The interior of these cells is separated from the cells' surroundings by the cell membranes, which are depicted as strong edges. So, if the cells are marked correctly, for example, using the determined cell markers, the watershed transform should be able to segment the cells properly. Unfortunately, closed cell membranes are required, since otherwise the water from different catchment basins would merge at locations that do not correspond to cell boundaries. Nonetheless, the watershed transform was applied to solve the task at hand. The results are introduced in Section 5.3.7.

In contrast to the approach described above, the watershed transform of cell images is often performed in a different way. Instead of employing the gradient magnitude image, the distance transform of a thresholded image, which for instance shows stained nuclei, is computed (cf. Section 5.2.2) [134][267]. Then, the watershed transform is applied to the distance map. So, clusters of cells can be separated. But this separation is only based on the cell clusters' shapes and not on the underlying image structures. Therefore, the resulting segments do not necessarily reflect the real shape of the cells under analysis.

The application of active contours solves several of the problems known from SRG and the watershed transform, since they incorporate edge information and are able to produce closed contours independent of the continuity of the visible edges. With respect to their computation, two techniques must be distinguished: parametric and geometric active contours.

Parametric active contours, which are often called *snakes* as well [111], deform an initial contour in such a way that it best fits the edges in its proximity. Therefore, it is necessary to select an initial contour which is near the real edges [25]. Then, snakes yield smooth and accurate results.

The topology of the curve is fixed. A snake $c(s)$ itself is represented by a parametric curve with arc length s (see Equation 5.13). Depending on the parametrisation of $c(s)$, the results might change, even if the initial contour remains equal [10, Chapter 4].

$$c(s) = \begin{pmatrix} x(s) \\ y(s) \end{pmatrix} \tag{5.13}$$

The optimisation of the snake is realised by minimising the energy functional $E^*_{\text{snake}}\big(c(s)\big)$. It is composed from several energy terms that assess the snake's shape with respect to the task at hand. The basic energy functional proposed by Michael Kass and his co-researchers is shown in Equation 5.14.

$$E^*_{\text{snake}}\big(c(s)\big) = \int_0^1 E_{\text{int}}\big(c(s)\big) + E_{\text{ext}}\big(c(s)\big)ds \tag{5.14}$$

$E_{\text{int}}\big(c(s)\big)$ describes the internal properties of the snake such as its continuity and curvature. In contrast, $E_{\text{ext}}\big(c(s)\big)$ incorporates external information like constraints interactively imposed by human users or knowledge extracted from the image under consideration. Here, the image intensity can be used directly, but more often knowledge on the gradients' magnitudes is applied.

In contrast to parametric active contours, *geometric active contours* are not represented by a contour c(s) explicitly. Instead, the contours evolve based on intrinsic geometric measures of an image [26]. The degree of evolution, which depends on the time t, is specified by a partial differential equation. Equation 5.15 gives an example, which was introduced by Vicent Caselles [26]. Here, ∇ represents the gradient operator, $\langle \underline{a}, \underline{b} \rangle$ the dot product of two vectors $\underline{a}$ and $\underline{b}$, and $|.|_2$ the Euclidean norm. Furthermore, $\underline{n}$ denotes the unit inward normal, κ the Euclidean curvature and g an edge detector function.

$$\frac{\partial c(s,t)}{\partial t} = \left(\kappa g\left(\left| \nabla I\big(c(s,t)\big) \right|_2 \right) - \left\langle \nabla g\left(\left| \nabla I\big(c(s,t)\big) \right|_2 \right), \underline{n} \right\rangle \right) \underline{n} \tag{5.15}$$

The evolution of a contour according to Equation 5.15 is equivalent to minimising the functional $E^*\big(c(s)\big)$, which is shown in Equation 5.16. $E^*\big(c(s)\big)$ resembles the snake energy functional $E^*_{\text{snake}}\big(c(s)\big)$, since it encompasses image-related information as well as a term assessing the shape of the contour. Here, γ denotes a constant.

$$E^*\big(c(s)\big) = 2\sqrt{\gamma} \int_0^1 g\left(\left| \nabla I\big(c(s)\big) \right|_2 \right) \left| c'(s) \right|_2 ds \tag{5.16}$$

A geometric active contour can be computed by means of level set approaches [197]. These, level set methods enable the efficient computation of propagating interfaces and boundaries. Here, an active contour is represented as the zero level set of a higher-dimensional function, i.e. the pixels where the function equals zero (cf. Figure 5.8).

This kind of implicit contour definition exhibits several features that are different to snakes. Firstly, its topology is not fixed. So, one geometric active contour can segment the complete image. Secondly, the computational effort is increased, as all pixels the contour can move to must be analysed. For snakes, a consideration of polygon points representing the contour is sufficient. The result of the application of a geometric active contour method [10, Chapter 4] to a bright-field image of Sf9 cells is depicted in Figure 5.9.

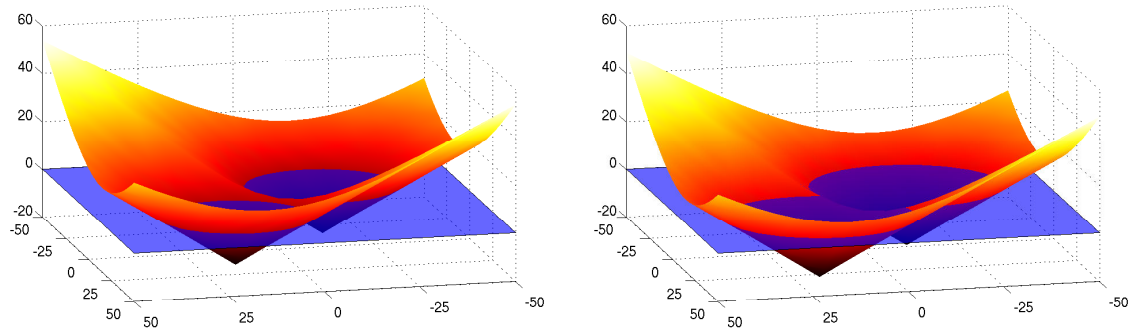


Figure 5.8: *Principle of level set methods.* A curve, for example segment boundaries, is regarded as the zero level of a higher-dimensional function. By slight variations of this function even the topology of the considered curve can be modified.

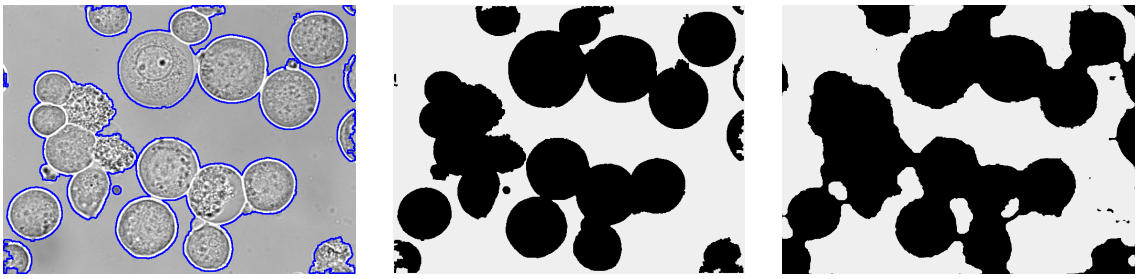


Figure 5.9: *Cell segmentation using a geometric active contour.* The whole image is segmented by one contour although multiple segments are created (a). Unfortunately, cell clusters are not split up properly. A binary visualisation (b) reveals strong similarities to the background image (c) that is used for the determination of cell markers.

Figure 5.9 shows that the ability to change the contour's topology might be a drawback with respect to the cell segmentation task at hand, as here the topology of the segments in question is known. This knowledge cannot be incorporated into a geometric active contour easily. Recently, several approaches enabling the usage of shape models have been proposed [67][92][277]. These techniques are usually applied to images showing single objects which match the model. If they were applied to the automatic segmentation of cells, each potential cell would have to be processed separately – similar to snakes. Then, they would further necessitate the selection of a suitable region of interest which corresponds to the applied model.

In principle, the usage of geometric active contours is very time consuming, as each pixel, the contour is able to move to, must be regarded in each iteration step. Therefore, the segmentation of the image shown in Figure 5.9 took more than 9 minutes on an AMD Athlon 64 processor (2GHz, 32-bit mode) although the image was downscaled from 1344×1024 pixels to 672×512 pixels. Admittedly, there are possibilities for optimisations. Nevertheless, the required computation time would be too long, especially, if high-throughput processing is the goal.

On the whole, the application of snakes appeared the most promising. In contrast to SRG, the watershed transform, and geometric active contours, they enable a fast independent segmentation of cells which might even overlap. Therefore, the formation of incorrect segments does not influence the result of adjoining regions. But they have further advantages. Firstly, they always yield closed contours even if the corresponding cell membrane is barely visible. Secondly, they enable the inclusion of context-specific knowledge such as membrane curvature and cell size without

using a complex model. So, the robustness can be improved. Finally, the computational effort for their computation is relatively low. Therefore, they were selected for the segmentation of Sf9 cells in bright-field images.

5.3.6 Parametric Active Contours

Several approaches have been proposed for the computation of parametric active contours, e.g. variational calculus [111], dynamic programming [7] and greedy methods [260]. I decided to focus on greedy approaches due to their efficient computability, stability and flexibility [234]. Nevertheless, a method based on variational calculus was investigated for comparison reasons.

Since the proposed approach aims at complete independence from user interactions while processing images, special requirements have to be fulfilled. In particular, the snake computation had to be modified so that, in contrast to the original approach, initialisations do not need to be a close approximation of the result. Then, an extension of the snake starting from the determined cell markers could be realised.

In principle, the application of gradient vector flow snakes [268] might solve this problem, as they increase the robustness of the basic snake approach. Unfortunately, they are based on variational calculus and diffusion equations, which results in a higher computational effort. Additionally, they are more sensitive to intracellular gradients.

Alternatively, Laurent D. Cohen [43] proposed a method to realise the growth of snakes by introducing an inflation force. This technique applies normal vectors of the contour in order to determine the direction of extension. As a result, the contour might overlap with itself if it is initialised with a concave cell marker. Hence, an alternative basis for the growth of the contour was utilised – the distance to the corresponding cell marker, which represents the initial contour. Equation 5.17 shows the employed energy functional $E^*_{\text{snake}}\big(c(s)\big)$, which assesses a snake. During the segmentation, E^*_{snake} is minimised.

$$\begin{aligned} E^*_{\text{snake}}\big(c(s)\big) \;=\; & \int_0^1 \Big[\alpha E_{\text{cont}}\big(c(s)\big) + \beta E_{\text{curv}}\big(c(s)\big) \\ & + \gamma\Big(E_{\text{dist}}\big(c(s)\big)\Big) E_{\text{ao}}\big(c(s)\big) \\ & + \delta\Big(E_{\text{dist}}\big(c(s)\big)\Big) E_{\text{dist}}\big(c(s)\big)\Big] ds \end{aligned} \tag{5.17}$$

The snake energy E^*_{snake} comprises the internal energies E_{cont} and E_{curv} as well as the external energies E_{ao} and E_{dist}. E_{cont} and E_{curv} control the continuity and curvature, respectively. Moreover, E_{cont} fosters equal spacing between neighbouring points [260]. E_{ao} represents the cell membrane image resulting from the algebraic opening (see Section 5.3.3) which has been inverted since the energies are to be minimised. E_{dist} reflects the distance from the initial contour computed by the distance transformation. Due to the growing of the contour, the distance has to be inverted, as well. Thus, a maximal considered distance Δ_{max} is required. It is set to the maximal cell radius increased by a tolerance interval of 20% (193 pixels in total).

The parameters α, β, γ, and δ control the influence of the respective energy terms. Here, γ and δ are modified with respect to E_{dist}. Their base values are denoted γ_0 and δ_0, respectively.

$$\gamma\Big(E_{\text{dist}}\big(c(s)\big)\Big) = \gamma_0 \cdot \frac{\Delta_{\max} - E_{\text{dist}}\big(c(s)\big)}{\Delta_{\max}} \tag{5.18}$$

$$\delta\Big(E_{\text{dist}}\big(c(s)\big)\Big) = \delta_0 + \gamma_0 - \gamma\Big(E_{\text{dist}}\big(c(s)\big)\Big) \tag{5.19}$$

According to Equation 5.18, $\gamma\big(E_{\text{dist}}\big)$ yields high values if E_{dist} is small, i.e. if the snake has a great distance to its initialisation. In contrast, high pixel values near the cell markers within the cells are suppressed. Equation 5.19 ensures that the sum of $\gamma\big(E_{\text{dist}}\big)$ and $\delta\big(E_{\text{dist}}\big)$ equals the sum of its base values γ_0 and δ_0, respectively. So, the extending force is reduced if the snake reaches a distance from its cell marker where the probability of membrane pixels is high. This mechanism is intended to prevent a premature abort of the growing procedure due to intracellular objects. Additionally, background pixels receive a high value of E_{dist} in order to avoid an extension of the snake in this region.

As mentioned above, two different implementations of active contours were applied within the scope of this thesis: a greedy approach (*greedy snakes*) and a method based on variational calculus (*vc-snakes*). Both of them employ the energy functional $E^*_{\text{snake}}\big(c(s)\big)$ shown in Equation 5.17. But they differ in the computation of the internal energies.

The greedy approach moves each snake point based on local considerations. Here, a point's neighbours along the contour as well as the external energies of potential destination points in its proximity are regarded: in particular, a window of 11×11 pixels centred at the respective snake point is utilised. Smaller windows do not allow growing if there are high external energies close to the cell marker and higher values enable the contour to jump over its cell marker or relevant edges. The external energies of the greedy snakes are computed according to Equation 5.20

$$E_{\text{ext}}\big(c(s)\big) = \gamma\Big(E_{\text{dist}}\big(c(s)\big)\Big)E_{\text{ao}}\big(c(s)\big) + \delta\Big(E_{\text{dist}}\big(c(s)\big)\Big)E_{\text{dist}}\big(c(s)\big) \tag{5.20}$$

The pseudocode of a greedy algorithm for snake computation is given in [260]. The only difference to the proposed technique lies in the computation of the external energies. So it can be transferred directly.

Although the optimisation is solely based on local considerations, the snakes might reach a state in which no further motions happen. Then the algorithm is terminated. Unfortunately, the snakes might oscillate, as well. These oscillations barely affect the quality of the segmentation. Therefore, a maximum number of iterations is required. This number is set to $\frac{1}{2}\Delta_{\max}$. As each snake point is able to bridge a distance of up to five points[1] in one iteration, the correct segmentation of large cells can be guaranteed by that.

In contrast to the greedy approach, the method based on variational calculus uses the formulas shown in Equations 5.21 and 5.22 for the adaptation of snake points [111]. Derivatives were approximated by finite differences. The contour $c(s)$ is represented by means of two vectors $\underline{x}(t)$ and $\underline{y}(t)$, which contain the snake's points at different iterations t.

[1] half of the considered neighbourhood's width

$$\underline{x}(t) = \left(\underline{A} + \eta \underline{I}\right)^{-1} \left(\eta \underline{x}(t-1) - \frac{\partial E_{\text{ext}}\left(\underline{x}(t-1), \underline{y}(t-1)\right)}{\partial \underline{x}}\right) \tag{5.21}$$

$$\underline{y}(t) = \left(\underline{A} + \eta \underline{I}\right)^{-1} \left(\eta \underline{y}(t-1) - \frac{\partial E_{\text{ext}}\left(\underline{x}(t-1), \underline{y}(t-1)\right)}{\partial \underline{y}}\right) \tag{5.22}$$

The parameter η, which controls the speed of the algorithm, is set to 0.25, since visual inspection showed that this setting leads to appropriate results; i.e., the cells can be segmented correctly. Higher values retard the growth of the snakes, whereas lower values cause strong fluctuations of the snake if it is situated on cell membranes. These fluctuations originate from the usage of the external energies' gradients, which lead to strong variations in the proximity of edges. Consequently, the snake does not converge, as might happen in the case of the greedy snakes. So, the segmentation of a cell must be stopped after a predefined number of iterations. Again, this number is chosen in such a way that all cells can be segmented properly, even if they exhibit the maximum radius. Nevertheless, the number of required iterations (2000) is significantly higher. The matrix $\underline{A}$ is symmetric, pentadiagonal and positive definite. Therefore, fast methods can be applied to the computation of the matrix inversion.

A comprehensive explanation of this variational calculus-based technique is given in [126]. It enables a more global consideration of the snakes than the greedy approach, which results in the ability to enforce the continuity and curvature constraints more strictly. In order to employ this method, again, only the computation of the external energies had to be adapted.

5.3.7 Results and Conclusion

In order to assess the segmentation results, 759 images of single cells, which had been extracted from 59 bright-field microscope images by a biological expert, were utilised. Figure 5.10 shows an example for such an extracted cell. This type of image is referred to as a *cell mask*.

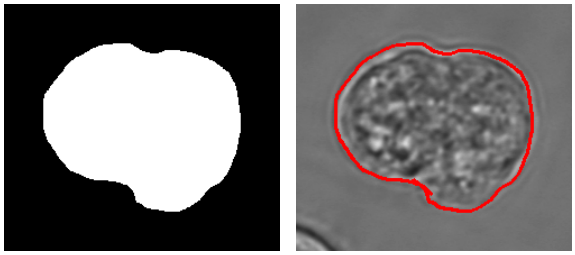


Figure 5.10: *Cell mask.* The left image shows an exemplary cell mask, which depicts all pixels of the corresponding cell in white and everything else in black. This cell was extracted from a bright-field image by a biological expert. It is shown by the image on the right. Here, the cell's contour, which was determined by the expert, has been drawn in red.

At first, the parameters required for the computation of the cell markers were determined. Applying the method suggested in Appendix A.1.1, the extension of the image background was performed using τ_{bg}=0.3 and n_{bg}=9. The length of the linear structuring elements employed during the detection of probable membrane pixels was set to l_{opt}=31 (cf. Appendix A.2).

Based on these values, 741 from 759 cells could be localised automatically, i.e. a corresponding cell marker was determined. So, 97.6% of the cells under analysis were available for evaluating the segmentation. The total number of determined cell markers amounts to 2,422, which appears very high. But it must be taken into account that not all cells, which are shown in the 59 regarded images, were extracted manually. Cells that are situated in the proximity of the image border

or are only partially visible, for example, had to be neglected. They are not suitable for protein localisation, since they do not allow for a consideration of whole cells. Even if they seem to be completely visible in a bright-field image, too small a distance to the image border might impair the localisation of proteins in certain cell compartments such as the cell membrane. Besides additional cells that were not manually extracted, the number of determined cell markers is increased, since very structured cells might receive more than one marker.

During the evaluation, the proposed snake algorithms were contrasted with the watershed transform (cf. Section 5.3.5). For the watershed transform, the implementation suggested in [206, Chapter 9] was utilised.

In order to assess the quality of the resulting segmentations, each segment j was compared to the corresponding cell mask by computing two error measures: the number A_j^{diff} of differently segmented pixels and the maximal distance d_j^{max} of both contours. These values measure the similarity of the comprised areas and the corresponding shapes, respectively. Furthermore, A_j^{diff} constitutes a well-known distance measure, which is called *Hamming distance* in order to honour its inventor, the American mathematician Richard Wesley Hamming. Hamming applied it so as to quantitate the similarity of codes in 1950 [81].

Both error measures are divided by the cell mask's size represented by its area A_j^{man} and the length a_j of the semimajor axis of the cell mask's approximation by an ellipse [65], respectively (see equations 5.23 and 5.24). So, the results of different cells become comparable.

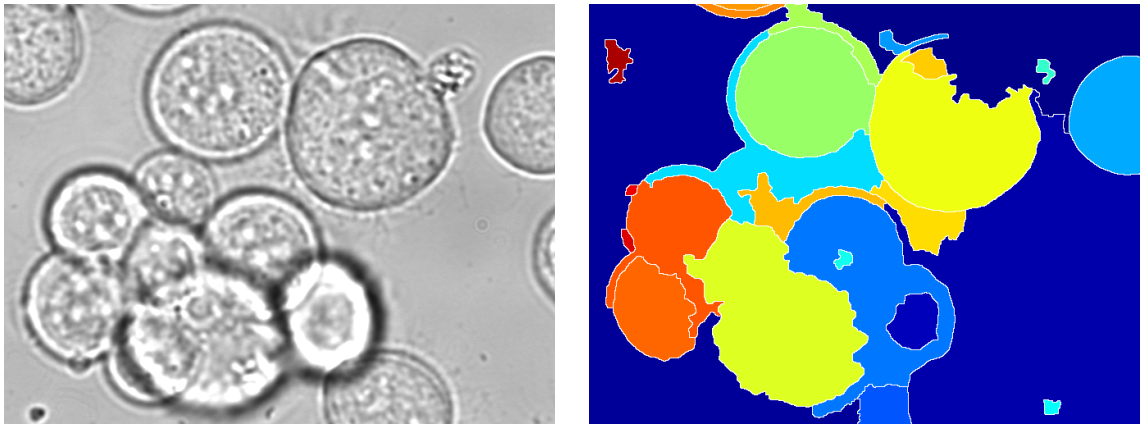
$$A_j = \frac{A_j^{\text{diff}}}{A_j^{\text{man}}} \tag{5.23}$$

$$d_j = \frac{d_j^{\text{max}}}{2a_j} \tag{5.24}$$

An optimal segmentation is characterised by $A_j{=}0$ and $d_j{=}0$. In contrast, a value of one indicates a segmentation error, which is as large as the whole cell – in terms of its area and its diameter, respectively. Depending on the applied segmentation procedure, even higher errors are possible.

In contrast to A_j, which is based on the number of incorrectly segmented pixels, d_j allows for the consideration of the segments' shapes. Even if A_j remains constant, d_j varies depending on the position of incorrectly segmented pixels with respect to the reference segment. The behaviour is required, as some segments that differ from the cell mask only in a relatively small number of pixels extend to image regions which possess a large distance to the reference segment. Such segments may occur especially in the case of the watershed transform (see Figure 5.11). In contrast, active contours impose constraints that facilitate the formation of segments reflecting the real shape of Sf9 cells.

By virtue of the characteristics of the regarded segmentation problem, a more complex evaluation procedure, such as the one proposed by Jayaram K. Udupa and his co-researches [237], was not necessary. So, for example, all pixels belonging to a cell are assumed to be equally important and no operator needs to be trained, as the final segmentation procedure is fully automated. However, the three main aspects, the evaluation of which Udupa and his colleagues demanded, are covered. These aspects are reproducibility, accuracy and efficiency. Using the proposed error measures A_j and d_j, a comparison of the considered methods in terms of reproducibility and accuracy became possible. The efficiency or rather the required processing time is given separately.



(a) (b)

Figure 5.11: *Segments resulting from the watershed transform.* A bright-field image (a) was segmented by means of the watershed transform. Each segment received a different colour and pixels belonging to watershed lines were dyed white (b). A comparison of both images reveals that several largely correct segments exhibit narrow extensions into the area of adjoining cells.

The watershed transform was applied directly to all 59 images, as it does not require any additional parameters for the segmentation. In contrast, the energy weights must be chosen in order to apply the snake approaches. This was achieved by image-wise cross-validation; 58 images served as the basis for the determination of the weights and the remaining image was employed for testing.

With respect to the greedy snakes, only the weights β, γ_0, and δ_0 had to be set. For α the arbitrary value of 0.5 was selected, since only the relation of the weights is relevant if the greedy snakes are applied. For β, γ_0, and δ_0 an exhaustive search within the interval $]0, 2]$ was performed. Here, a step size of 0.05 was chosen, which resulted in 41^3=$64,000$ combinations which had to be evaluated. Both A_j and d_j were determined by employing the respective optimal weights, i.e. the weights that minimised the error. Using an AMD Athlon 64 processor operating at 2GHz, the optimisation of the energy weights with respect to a single bright-field image (1344×1024 pixels) takes about one day.

Unfortunately, not all parameters, which have been determined using the greedy approach, can be transferred to the vc-snakes, as here α and β are incorporated differently. Nevertheless, γ_0 and δ_0 can be adopted. So, an additional exhaustive search was performed for α and β. Here the interval $[0.0001, 0.04]$ and a step size of 0.0005 were utilised. Therefore, 80^2=$6,400$ combinations of weights had to be considered. As in the case of the greedy snakes, the optimisation of the energy weights regarding a single bright-field image takes approximately a whole day using an AMD Athlon 64 processor (2GHz, 32-bit mode).

Although the chosen intervals for the energy weights α and β differ significantly from the ones of the greedy snakes, they cover the complete range for which a correct segmentation was observed. The transfer of γ_0 and δ_0 is necessary, as otherwise an exhaustive search for all weights would not be possible in an acceptable period of time; a complete evaluation using the proposed intervals for all four weights leads to 80^2+41^2=$10,758,400$ different combinations to be utilised.

Besides the relevant parameters, the parametrisation of a snake, i.e. its representation, is crucial for the segmentation. As indicated in Section 5.3.4, the initial snakes are obtained by means of a contour following algorithm [103, Chapter 9] applied to the cell markers. Based on this initial

polygon, the snakes are extended. As a result, the distance between neighbouring snake points is increasing. Therefore, after each iteration it is checked whether additional points need to be inserted (see Appendix A.3). Furthermore, the vc-snakes tend to concentrate points at certain locations, for example, at the image border. So new points are inserted continuously. In order to prevent the vc-snakes from using too high a number of points, polygon points are removed if they represent the same pixel.

Table 5.1 shows the mean μ, the standard deviation σ and the median m of the segmentation errors A_j and d_j caused by the three considered segmentation techniques. Here, the median was computed by sorting the respective segmentation errors for all 741 cells and selecting the element at position 371. Therefore, 50% of the considered cells were segmented with a smaller or at least equal error.

	A_j			d_j		
segmentation method	μ	σ	m	μ	σ	m
watershed transform	11.070	20.522	0.251	1.802	2.779	0.210
greedy snakes	0.174	0.118	0.147	0.110	0.079	0.093
vc-snakes	0.189	0.130	0.160	0.122	0.078	0.105

Table 5.1: *Segmentation errors.* In order to evaluate the segmentation results, the mean μ, the standard deviation σ and the median m of both segmentation errors are considered.

Both snake approaches performed the segmentation with small errors. Here, the segmentation using the greedy snakes yields a slightly more accurate segmentation. The bad results of the watershed transform with respect to mean and standard deviation originate in a number of segments that do not resemble the shape of single cells at all. They rather cover fractions of the image background and other cells. This is a result of gaps in the cell membranes that enable a segment to extend to other image regions, which, as a consequence, cannot be segmented properly as well. By virtue of the large differences to the respective cell masks, which are determined based on the corresponding cell marker, μ and σ of both error measures are increased. Nevertheless, the median m shows that at least 50% of the considered cells are significantly better segmented than the average. The applied active contours do not suffer from such gaps, as they incorporate prior knowledge on the cell shape. Furthermore, an incorrect snake does not affect the segmentation of other cells, since the snakes are independent of each other.

In order to illustrate the results shown in Table 5.1, Figure 5.12 depicts the segmentation of a bright-field image using the greedy snakes. The applied image was not used for any optimisations of parameters. Nevertheless, the segmentation of almost all cells is correct. Here, partially visible cells and cells, which are situated close to the border, are negligible, since they cannot be employed for protein localisation. However, the localisation yields cell markers that do not correspond to cells. As a result, some snakes also do not represent cells. These segments have to be sorted out (see Section 5.4).

On average, A_j is higher than d_j for all considered techniques, which results from the fact that A_j is an area measure while d_j reflects a distance. A_j rises depending on the squared distance from a segment to its reference contour. As a result, the difference between A_j and d_j grows if the distance to the respective cell mask increased. This results in an excessive rise of the mean and the standard deviation of A_j with respect to the watershed transform that can be observed in Table 5.1. In contrast, the median differs only slightly. Since d_j assesses not only the number

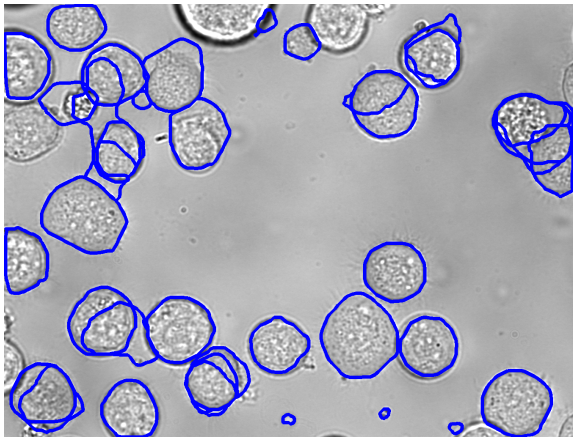


Figure 5.12: *Segmentation using the greedy snakes.* The depicted test image was segmented properly by means of the greedy snakes, which are visualised as blue contours. But, although the majority of the visible cells was segmented correctly, there are segments that do not correspond to cells.

of correctly segmented pixels but the resulting shapes, as well, it reflects the quality of the segmentation more appropriately than A_j. Therefore, mainly d_j was applied within the scope of the following investigations.

Besides the segmentation errors, the computation time was measured using an AMD Athlon 64 processor (2GHz, 32-bit mode) like in the previous experiments: The optimal energy weights were determined for all 59 images and the error d_j.[2] Since all three methods use the same cell markers, the time required for localising the cells, which amounts to approximately 4.5s, was subtracted. The greedy snakes achieved the best result. They segmented an image in 2.2s on average. The watershed transform took about 15.6s and the vc-snakes required 84.3s per image. This relatively bad result of the vc-snakes partly originates from the reparametrisation procedure (see Appendix A.3). Since in comparison to the greedy snakes more iterations are required, resampling is performed more often. But as the snake moves considerably slower, a check whether new points need to be inserted is not required after each iteration. Therefore, the vc-snakes were evaluated again. But now resampling was performed after every 20th iteration and the number of iterations was decreased from 2000 to 1920. So, the number of resampling steps performed by the greedy snakes and the vc-snakes is comparable ($\frac{1}{2}\Delta_{\max}\approx 96$). As a result, the images were segmented in 9.2s on average. However, this is considerably slower than the greedy snakes.

Although the greedy snakes enable a very fast segmentation, the computational effort could be further reduced if the number of iterations was decreased (see Appendix C.1). Moreover, it would be possible to reduce the size of the considered images. As I successfully applied the greedy snakes to images which were taken at 40× instead of 60× magnification [232], a slight reduction should not cause any additional problems. However, the contrast of cell membranes might be impaired if the image size is decreased too much.

5.4 Rejection of Non-Cell Segments

Section 5.3.7 has shown that the segmentation yields segments which depict single cells. But segments containing other images structures are also created. This results from the fact that candidate cells determined during the localisation do not necessarily correspond to real cells. They may be

[2] The optimisation of the greedy snakes led to the following weights: α=0.5, β=1.7, γ_0=0.4 and δ_0=0.7. As γ_0 and δ_0 can be transferred to the approach based on variational calculus, only α and β had to be computed again. So, the vc-snakes employed α=0.0121 and β=0.0041.

caused by noise or image objects such as dirt as well. Furthermore, some cells are segmented incorrectly; that is, with large errors. These segments should be rejected. In contrast to the cell markers, the complete segments allow for a comprehensive characterisation that enables the rejection of these non-cell segments.

Such a characterisation was realised by means of adequate features, which are introduced in Section 5.4.1. They were applied to train classifiers. Hence, appropriate training and validation datasets had to be acquired. This was achieved by utilising the 759 cell masks, which were employed during the evaluation of the localisation and segmentation techniques (cf. Section 5.3.7). Additionally, non-cell samples were collected using the same 59 bright-field images that the cell masks originate from. The process of dataset generation is addressed in Section 5.4.2.

But not all of the considered features are useful with respect to the task at hand. Therefore, unnecessary features should be omitted. Here, the performance of the classifiers in terms of computational load and required training samples can and should be decreased. Using the generated dataset, such an evaluation and reduction of the basic feature set was performed (see Section 5.4.3).

Based on the resulting feature sets, a rejection of non-cell samples can be carried out. But depending on the type of classifier used, the results and prerequisites vary considerably. Therefore, Section 5.4.4 introduces and compares several classification methods.

Finally, the classification methods as well as the feature reduction methods are evaluated (see Section 5.4.5). Here, a technique will be selected that is suited to the task of protein localisation, which is to be performed after the cells have been recognised.

5.4.1 Determination of Adequate Features

The applied features are based on information gained during the segmentation. So, the effort required for their computation can be reduced. The basic feature set comprises simple shape features such as the area and the perimeter of the image segments. In addition, histogram-based features, essentially statistical moments and quantiles, are computed on the original image, the image containing possible membrane pixels and its point-wise square, as well as the image depicting the background. Here, it is particularly important to consider cell membrane pixels, since the intracellular image features are similar for different cells. As the position of these membrane pixels within a segment is relevant, three segment-specific image regions are examined in addition to the segment itself (see Figure 5.13).

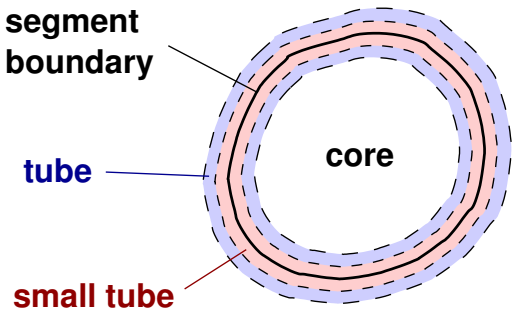


Figure 5.13: *Considered image regions.* Besides a segment, two tubes around its boundary and its core are employed for computing appropriate features that allow for a separation of cell and non-cell regions.

The first region constitutes a tube around the segment contour (radius: 10% of the mean cell diameter). There, the fraction of potential membrane points should be high. Furthermore, a small tube with a radius of 5% of the mean cell diameter is considered as well, in order to detect very close membrane pixels. The third region constitutes the core of a segment, i.e. the part that is not

covered by any tube. No membrane pixels should be situated here. On the whole, a total of 111 features is employed (see Appendix B.1 for a more detailed description). Except for the segment area and perimeter, all these features are invariant with respect to rotation and scaling. This is a crucial property, since the cells vary in diameter and orientation.

5.4.2 Generation of Datasets

In order to acquire a reasonable amount of training data, three different methods were applied. Firstly, the available cell masks were utilised. Secondly, the corresponding bright-field images were segmented automatically using the greedy snakes. Here, a large set of different values was utilised for the relevant parameters, in particular, the energy weights of the active contours. So, a multitude of segments was generated and compared with the corresponding cell masks on the basis of the error measures A_j and d_j (see Section 5.3.7). Small values of d_j ensure that the respective segments resemble the shape of the cell masks. In contrast, segments exhibiting a small error A_j might posses a completely different shape. Therefore, the maximum of d_j and A_j was applied as an alternative to d_j rather than A_j itself. Thus, the following two similarity measures E_j^{seg} were utilised for the generation of additional training samples:

$$E_j^{\text{seg}} = d_j \tag{5.25}$$

$$E_j^{\text{seg}} = \max\left(A_j, d_j\right) \tag{5.26}$$

In order to increase the training set, segments j with $E_j^{\text{seg}}<0.1$ are considered as cells as well. This can be justified by the fact that the boundary of cells is difficult to define exactly (cf. Appendix C.2) and the mean segmentation error amounts to approximately 0.1 if the greedy snakes are applied (cf. Section 5.3.7). Furthermore, segments leading to $E_j^{\text{seg}}\geq 0.33$ are assumed to represent non-cell segments.

Eventually, further non-cell segments were selected from automatically segmented images by hand. In order to enable the generation of a high number of non-cell segments, the localisation procedure was slightly modified. So 4,557 rather than 2,422 cell markers were determined using the 59 bright-field images under analysis (cf. Section 5.3.7).

On the whole, the training set encompasses automatically generated patterns and manually determined examples. Therefore, the final classifier is not only adapted to the cell masks, but to the segmentation procedure as well.

5.4.3 Feature Reduction

Feature reduction is a beneficial preprocessing step for classification tasks, in particular, due to the decreased dimensionality of the data resulting in a higher performance of a classifier. Here, it has to be distinguished between methods that remove irrelevant features and techniques removing redundant features. Both of them are important means for increasing the quality of data [170].

With respect to feature reduction, two principal methods must be distinguished: *feature extraction* and *feature selection* [129][170]. The first one performs a functional mapping of existing features, while the latter selects a suitable subset. Here the advantage of feature selection methods consists in the fact that rejected features do not need to be computed. In contrast, a feature extraction technique requires all features to be calculated and combines them appropriately.

Feature selection methods can be further divided into *filters* and *wrappers* [129]. While a filter is applied before the classification and operates independently of it, a wrapper optimises the actual classification results. Therefore, wrapper approaches usually outperform filters. Unfortunately, they increase the computational load as well.

In order to account for the high number of available training samples (cf. Table 5.2), I decided to resort to methods which are efficiently computable, in particular, filters and feature extraction techniques. Hence, I analysed four different methods: an adapted correlation analysis, principal component analysis, independent component analysis and kernel principal component analysis. They are introduced below, followed by a brief summary.

Correlation Analysis

As mentioned above, the first method used for reducing the employed feature set consists in *correlation analysis* (CA) – a filter approach. The theoretical foundations of correlation in general were laid by the British polymath Sir Francis Galton at the end of the 19th century [69][86]. Here, the correlation r_i [82, Chapter 7] between the similarity criterion E_j^{seg} of all segments j and the features x_i is analysed according to Equation 5.27. $\bar{E}^{\text{seg}}$ and $\bar{x}_i$ denote the respective mean values and j the index of a segment from the training set. By this procedure, only correlated features with $\mid r_i \mid > 0.2$ are preserved.

$$r_i = \frac{\sum_j \left(x_{ij} - \bar{x}_i\right)\left(E_j^{\text{seg}} - \bar{E}^{\text{seg}}\right)}{\sqrt{\sum_j \left(x_{ij} - \bar{x}_i\right)^2 \sum_j \left(E_j^{\text{seg}} - \bar{E}^{\text{seg}}\right)^2}} \tag{5.27}$$

According to Equation 5.27, only linear correlations are considered [57, Chapter 4]. Therefore, the threshold for rejecting features was chosen to be rather low so as to prevent features exhibiting a small linear correlation from being removed.

Unfortunately, the manually determined non-cells must be neglected, since for them no value for E_j^{seg} exists, as they cannot be associated with a cell mask. In order to incorporate them into the selection process, nonetheless, they receive the maximal E_j^{seg} that has occurred and r_i is computed again. Features whose r_i has changed by more than 20% of the maximal difference that occurred are considered important for the recognition of the manually selected non-cells and retained as well.

The measure r_i was chosen, as it enables a simple and fast rejection of correlated features using numerical values. But in principle, r_i could be substituted with more complex correlation measures enabling the investigation of nonlinear relationships, for instance mutual information [12], if required.

Principal Component Analysis

Besides correlation analysis, *principal component analysis* (PCA) [99, Chapter 6][218] was examined with respect to the task at hand. It constitutes a feature extraction method which removes redundancies within the regarded data. The foundations of the PCA were laid by the British mathematician Karl Pearson, who investigated the fitting of lines and planes to points in higher-dimensional spaces at the beginning of the 20th century [159]. If applied to the problem at hand, the PCA can be used to determine a new orthogonal coordinate system where the features are uncorrelated and the axes correspond to the directions in which the training data exhibit the strongest variances. A new feature y_j, also called *principal component*, is computed by a simple linear com-

bination of the centred feature vectors $\underline{x}$ and a transformation vector $\underline{e}_j$. Here, $\underline{e}_j$ is referred to as *principal axis*.

$$y_i = \underline{e}_j^T \underline{x} \tag{5.28}$$

If $\underline{y}$ denotes the transformed feature vector, Equation 5.28 can be rewritten in matrix form (see Equation 5.29). Here, $\underline{\underline{E}}$ symbolises the transformation matrix. While the feature vector $\underline{x}$ comprises p features, $\underline{y}$ is composed from q elements. Provided that q equals p, the original data can be reconstructed from $\underline{y}$ independent of their distribution.

$$\underline{y} = \underline{\underline{E}}\,\underline{x} \qquad \text{with} \qquad \underline{\underline{E}} = \begin{pmatrix} \underline{e}_1^T \\ \underline{e}_2^T \\ \vdots \\ \underline{e}_q^T \end{pmatrix} \tag{5.29}$$

In order to reduce the dimensionality of the feature space, q must be chosen smaller than p. Here, a limited loss of information is usually approved. The transformation vector $\underline{e}_1$ represents the direction of the strongest variance, $\underline{e}_2$ the direction of the second-strongest variance and so on. Hence, a maximum amount of variance from the original data is preserved in the transformed features. In addition, the loss of information can be quantitated in terms of the variance.

A PCA can be easily computed based on the covariance matrix $\underline{\underline{C}}_{\underline{x}}$ of the original features (see Equation 5.30). Here n denotes the number of available samples and $\underline{\mu}$ their mean vector.

$$\underline{\underline{C}}_{\underline{x}} = \frac{1}{n} \sum_{i=1}^{n} \left(\underline{x}_i - \underline{\mu}\right)\left(\underline{x}_i - \underline{\mu}\right)^T \tag{5.30}$$

The principal axes are the unit-length eigenvectors of $\underline{\underline{C}}_{\underline{x}}$. Their ordering depends on the corresponding eigenvalues λ_j which equal the variance of the new features y_j. Thus they satisfy: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$.

If the PCA is applied, two tasks have to be dealt with [218]. Firstly, a number $q \leq p$ of relevant eigenvectors must be chosen. Secondly, an adequate normalisation of the feature vectors should be performed, especially in the case that the features are measured at different scales.

The first problem can be solved in numerous ways. q can be selected based on the fraction of variance of the original data which is to be maintained or depending on the mean variance of the principal components. Furthermore, a visual inspection of the eigenvalues might reveal important information. I decided to ascertain q with respect to the amount of variance which should be maintained, since this enables an objective comparison of different data using numerical values rather than visual impressions. Within the scope of this thesis, the number of eigenvalues is usually chosen to maintain 99% of the variance of the original data.

Problems resulting from different scales of the regarded features are easily circumvented if the centred features are divided by their standard deviation. Otherwise, features at large scales would be incorporated exorbitantly.

Unfortunately, the amount of variance does not necessarily reflect the class separability. Therefore, the PCA might diminish the classification accuracy even if the total variance is only decreased slightly.

Independent Component Analysis

As a third feature reduction method, the *independent component analysis* (ICA) [99, Chapter 7] is employed. In fact, it does not reduce the number of features directly. This is rather a result of the preprocessing required to perform an ICA. In principle, the ICA performs a modification of the utilised coordinate axes. In contrast to the PCA, the ICA does not only select directions exhibiting a high variance of the patterns, but considers other properties such as the clustering structure, as well (cf. [99, Section 8.5]).

The goal of the ICA consists in the determination of independent sources which caused a set of observations $\underline{x}$. Within the scope of my thesis, these observations correspond to the feature vectors. If the ICA is applied, they are assumed to have been generated by means of a weighted sum of the sources. Using vector notation, this yields Equation 5.31, which constitutes the basic model of the ICA. Here, $\underline{s}$ denotes the vector of source signals and $\underline{A}$ the mixing matrix. The source signals are referred to as *independent components*.

$$\underline{x} = \underline{A}\,\underline{s} \tag{5.31}$$

The major problem of the ICA consists in the fact that the mixing coefficients as well as the source signals are unknown. In order to determine these quantities, only one assumption is made: It is supposed that the sources are mutually statistically independent. But there are two further restrictions with respect to the basic ICA model: Firstly, the source signals must have a nongaussian distribution and, secondly, the mixing matrix is assumed to be square.

Gaussian independent components are not allowed, as this kind of distribution is fully characterised by its first two moments – the mean and the variance. But the ICA relies on higher-order information. Hence, in the case of Gaussian mixtures, the mixing matrix cannot be inferred [99, Chapter 7]. Furthermore, uncorrelated Gaussian mixtures are always independent. So, the usage of the ICA is not necessitated here. Instead, a PCA could be performed.

The mixing matrix is assumed to be square in order to alleviate the computation, since then the independent components can be determined according to Equation 5.32. In principle, the number of signal mixtures (features) must be at least as high as the number of source signals [210, Chapter 2]. However, there might be more signal mixtures than sources. Then the number of signal mixtures has to be reduced.

$$\underline{s} = \underline{A}^{-1}\underline{x} \tag{5.32}$$

As a result, the number of independent components to be computed equals the number of mixtures, i.e. with respect to the task at hand, the number of features. If there are redundant mixtures, they can be omitted.

The results yielded by the ICA have two limitations: Firstly, the independent components' variances cannot be determined, as any scaling of the independent components could be made undone by means of dividing the corresponding row of $\underline{A}$ by the same factor. Therefore, the independent components are usually assumed to have unit variance. However, the sign of an independent component cannot be derived. Secondly, the order of the independent components is arbitrary. Therefore, any permutation of the independent components might result from an application of the ICA.

In practice, it has proven beneficial to perform an operation called *whitening* before computing an ICA [99, Section 7.4]. So the number of parameters to be estimated is diminished, which is es-

pecially important if high-dimensional data are applied. In comparison to the ICA, the whitening is a very simple linear transformation that can be computed quite efficiently by standard procedures. In principle, it constitutes a decorrelation and a scaling of the elements of the feature vector, which are carried out subsequently. The resulting new mixtures (features) z_j are uncorrelated and exhibit unit variance. They are computable by means of the eigenvector matrix $\underline{E}$ and the eigenvalues λ_j (see Equation 5.33), which can be obtained using the PCA.

$$\underline{z} = \underline{\Lambda}^{-\frac{1}{2}} \underline{E}\, \underline{x} \qquad \text{with} \qquad \underline{\Lambda}^{-\frac{1}{2}} = \begin{pmatrix} \lambda_1^{-\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & \lambda_2^{-\frac{1}{2}} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_p^{-\frac{1}{2}} \end{pmatrix} \tag{5.33}$$

With respect to $\underline{z}$, the basic model of the ICA (cf. Equation 5.31) is modified as shown by Equation 5.34.

$$\underline{z} = \underbrace{\underline{\Lambda}^{-\frac{1}{2}} \underline{E}\, \underline{A}}_{\underline{A}^*} \underline{s} \tag{5.34}$$

If whitening has been performed, the computation of the ICA boils down to the determination of an orthogonal transformation matrix $\underline{A}^*$. Then, only $p(p-1)/2$ free parameters have to be determined instead of p^2. Since within the scope of my thesis the dimensionalities of the considered feature vectors are rather high, the ICA is always computed using whitening as a preprocessing step. Furthermore, the dynamic ranges of the new features are adjusted by whitening [63, Chapter 9]. So all features are considered equally important, which is advantageous as well.

In order to enforce the property that the mixing matrix $\underline{A}$ is orthogonal, an explicit *orthogonalisation* method [99, Chapter 6] is applied. The classic Gram-Schmid orthogonalisation, which works sequentially, leads to an accumulation of errors depending on the dimensionality of the feature space. Therefore, I preferred a symmetric orthogonalisation technique that treats all vectors, contained in the mixing matrix $\underline{A}$, equally.

Assuming there are more signal mixtures (features) than sources, the number of mixtures must be reduced. Otherwise the mixing matrix would not be invertible (cf. Equation 5.32). If the correct number of sources is unknown, it must be estimated; or at least, redundant mixtures need to be removed. This can be achieved by means of the PCA during the whitening step [42][210, Chapter 2]. Jean-Pierre Nadal and his co-researches showed that projecting data on the space spanned by the q eigenvectors corresponding to the q largest eigenvalues maintains the q strongest independent sources [150]. But weak sources might be rejected similar to weak principal components. However, since the PCA is employed as a preprocessing step anyway, I decided to utilise it for the removal of redundant information as well.

The actual computation of the ICA was realised by means of the fixed-point algorithm proposed in [99, Section 8.4.3]. As fixed-point algorithms are a technique that performs a batch processing of a large amount of data in parallel, they are very fast [98]. Due to the high number of features regarded within the scope of my thesis, this constitutes an important quality.

Kernel Principal Component Analysis

The feature reduction approaches discussed so far do not account for nonlinear relationships of the data. Therefore, a modification of the basic PCA termed kernel principal component analysis,

or *kernel PCA* [193] in short, was taken into consideration. By virtue of its relationship to the PCA and the ICA, it seemed to be an excellent candidate for comparative analyses. But after performing preliminary experiments, the kernel PCA was dropped due to practical reasons, which will be detailed below.

The kernel PCA takes advantage of the so-called *kernel trick* [194, Chapter 1], which enables dot products in arbitrary algorithms to be substituted with an alternative similarity measure referred to as a *kernel* [194, Chapter 2]. In principle, such a kernel represents a dot product in an alternative, usually higher-dimensional space $\mathcal{H}$. The corresponding map from the original input space $\mathcal{X}$ to $\mathcal{H}$, which might be nonlinear, is denoted by $\Phi(\underline{x})$. So the computation of a kernel k using the vectors $\underline{x}_1$ and $\underline{x}_2$ can be performed according to Equation 5.35.

$$k(\underline{x}_1, \underline{x}_2) = \left\langle \Phi(\underline{x}_1), \Phi(\underline{x}_2) \right\rangle \tag{5.35}$$

The major advantage of using kernels in comparison to a direct transformation of the data consists in the fact that the map $\Phi(\underline{x})$ does not have to be computed explicitly. Examples for popular kernels as well as their properties are discussed in Section 5.4.4 and [194, Chapter 2].

Instead of regarding the covariance matrix $\underline{C}_{\underline{x}}$ known of the basic PCA (cf. Equation 5.30), the kernel PCA analyses the covariance matrix $\underline{C}_{\Phi(\underline{x})}$ of the data in the new feature space $\mathcal{H}$ (see Equation 5.36). In order to simplify the computations, the samples are assumed to be centred. The problem of centring will be discussed below.

$$\underline{C}_{\Phi(\underline{x})} = \frac{1}{n} \sum_{i=1}^{n} \Phi(\underline{x}_i) \Phi(\underline{x}_i)^T \tag{5.36}$$

As the map $\Phi(\underline{x})$ should not be computed explicitly, a *dual representation* of the eigenvectors $\underline{e}_j$ of the covariance matrix is constructed. The eigenvectors which have nonzero eigenvalues λ_j can then be determined according Equation 5.37. In particular, they are expressed as a weighted sum of the transformed input data. The respective weights are denoted by α_{ji}.

$$\underline{e}_j = \sum_{i=1}^{n} \alpha_{ji} \Phi(\underline{x}_i) \tag{5.37}$$

Using the dual representation, the weights can be obtained by solving the following eigenvalue problem (see Equation 5.38).

$$n\lambda_j \underline{\alpha}_j = \underline{K}\, \underline{\alpha}_j \tag{5.38}$$

Here, $\underline{K}$ denotes the *Gram matrix*, also called the *kernel matrix*. The Gram matrix encompasses the kernels for all pairwise combinations of the sample data.

$$\underline{K} = \begin{pmatrix} k(\underline{x}_1, \underline{x}_1) & \cdots & k(\underline{x}_1, \underline{x}_n) \\ \vdots & \ddots & \vdots \\ k(\underline{x}_n, \underline{x}_1) & \cdots & k(\underline{x}_n, \underline{x}_n) \end{pmatrix} \tag{5.39}$$

Like with the basic PCA, the principal components are determined by computing the projection of the data points on the principal axes, which need to be normalised before. However, these

operations refer to the space $\mathcal{H}$. In terms of kernels, the required projection is computable as shown in Equation 5.40.

$$\left\langle \underline{e}_j, \Phi\left(\underline{x}\right) \right\rangle = \sum_{i=1}^{n} \alpha_{ji} k\left(\underline{x}_i, \underline{x}\right) \tag{5.40}$$

The resulting properties are similar to the original PCA; in particular, the principal components are uncorrelated and represent the orthogonal directions of the highest variances. But they refer to $\Phi\left(\underline{x}\right)$ rather than $\underline{x}$. Here, the map $\Phi\left(\underline{x}\right)$ is implicitly determined by means of the selected kernel.

Up to now, one crucial problem has been left open – the centring. The PCA requires all data to be centred. Hence, the kernel PCA has the same requirement. But now, the data points must be centred in $\mathcal{H}$. With respect to the training data, this can be achieved by utilising Equation 5.41 [194, Chapter 14], which describes the computation of a new Gram matrix $\underline{\tilde{K}}$.

$$\underline{\tilde{K}} = \underline{K} - \left(\frac{1}{n} \cdot \underline{I}\right) \underline{K} - \underline{K} \left(\frac{1}{n} \cdot \underline{I}\right) + \left(\frac{1}{n} \cdot \underline{I}\right) \underline{K} \left(\frac{1}{n} \cdot \underline{I}\right) \tag{5.41}$$

In [217] a corresponding formula is introduced that enables the computation of a Gram matrix with respect to an available set of test data. Compared to the original PCA, the computational load for performing centring is significantly higher.

The computations described by Equation 5.38 and Equation 5.40 have three serious consequences: Firstly, the size of the Gram matrix depends on the number of training samples available; i.e., it might be very large. So, a training set comprising $100,000$ samples, which roughly corresponds to the size of the actually applied sets (cf. Section 5.4.5), results in a Gram matrix requiring 37.3GB of memory.[3] Secondly, the complete training set is necessitated so as to extract the principal components of new data during the application of the approach. Thirdly, the number of principal components might be significantly higher than the number of original features, which clearly contradicts the goal of feature reduction. On the whole, the application of the basic kernel PCA technique to the cell recognition task under consideration is not reasonable.

Recently, some incremental approaches to realising the kernel PCA have been proposed [39][118]. They avoid the practical difficulties resulting from computing and storing the required Gram matrix. However, the third problem remains: Using large datasets, the number of principal components to be extracted might be very high and even significantly exceed the number of the original features. Therefore, it was rejected as a means of feature reduction for the investigated cell recognition task.

But although the kernel PCA is not an appropriate means of feature extraction with respect to the task at hand, alternative methods might have been beneficial; for instance, the *locally linear embedding* (LLE) [184] appeared promising. Based on simple local considerations, it allows for the learning of the global structure of nonlinear manifolds. However, as the employed linear methods sufficed to extract adequate features regarding the cell recognition task under analysis (cf. Section 5.4.5), I did not examine any further nonlinear feature extraction methods.

Summary

According to the discussion above, three of the regarded feature reduction methods are applicable to the task at hand: one filter (CA) and two feature extraction techniques (PCA and ICA). In order to diminish the computational load as much as possible, I decided to combine them. Therefore, the

[3] Here the size of one kernel value $k\left(\underline{x}_i, \underline{x}_j\right)$ is assumed to be four bytes.

proposed approach comprises two successive reduction steps: correlation analysis (CA) in order to remove irrelevant features as well as PCA and ICA, respectively, to decrease redundancies between features. Since the CA represents a filter approach, the number of features to be considered by PCA and ICA is diminished. This is especially beneficial for a high-throughput application.

5.4.4 Classification

Based on the feature sets resulting from the feature reduction methods described in Section 5.4.3, the discrimination between cell segments and non-cell segments is performed. Here the overall task of protein localisation needs to be accounted for. In particular, a protein localisation method is aimed at, which enables the integration of new patterns during the application. Therefore, a classifier capable of on-line learning would be beneficial. Thus, I decided to use a special kind of neural network called ARTMAP, which is designed to enable fast and incremental learning. ARTMAP nets and its components have been applied to numerous problems in pattern recognition, e.g. the diagnosis of genetic abnormalities [247], vowel classification [261], solving sensorimotor tasks [226][227][230] and the analysis of electrocardiograms [212]. In addition, I applied support vector machines (SVMs). This type of learning architecture has been reported to achieve excellent classification results and a very good generalisation to unknown patterns; its application fields are the classification of proteins [243], the detection of computer crimes [273] and face recognition [124], for example. Here, SVMs are applied as a reference method and compared to the ARTMAP-based techniques. Both types of classification approaches are detailed in the following:

Simplified ARTMAP Networks

ARTMAP [22] comprises a family of supervised neural networks that were originally developed by Gail A. Carpenter and her co-researchers. The ARTMAP architecture consists of two unsupervised *adaptive resonance theory* (ART) networks. In principle, one of these modules clusters the input vectors and the other the output. They are combined via an associative learning network called *map field*, which enables a linking of corresponding input and output clusters.

The adaptive resonance theory was introduced as a model for information processing in the human brain by Stephen Grossberg at Boston University [78]. It suggests the existence of resonance processes, which occur if a match between neural activities is reached. Furthermore, it assumes that previously uncommitted nodes get involved in the process in the case of mismatch. Besides exploiting these mechanisms in order to develop innovative artificial networks, they can be applied to explain and predict numerous known psychological effects.

To date, there exists a great variety of ARTMAP networks, for instance, hypersphere ARTMAP [8], ellipsoid ARTMAP [9], Gaussian ARTMAP [261], distributed ARTMAP [24] and fuzzy ARTMAP [21]. All these networks share a common basic structure. But the actual processing of data might differ significantly.

One of these architectures, which is especially well-suited for fast, stable and incremental learning of numerical data, is the *fuzzy ARTMAP*. It has been shown to reach very accurate results after only one epoch of training if used in batch mode [247]. Using incremental training [247], the fuzzy ARTMAP has been reported to reach accuracies comparable to other well-known methods like multilayer perceptrons [202][275, Chapter 7] and support vector machines [194, Chapter 7][243].

Although the fuzzy ARTMAP takes advantage of the computational simplicity of the unsupervised *fuzzy ART* [23], its structure is rather complex, since it constitutes a seven-layered recurrent

neural network. So, the fuzzy ARTMAP structure was further simplified, which resulted in the *simplified fuzzy ARTMAP* (SFAM) [113][241]. In contrast to the original fuzzy ARTMAP, which has been proven to be a universal function approximator [246], the SFAM is exclusively designed for solving classification tasks; i.e., assigning discrete class labels to input vectors.

Like the original fuzzy ARTMAP, the SFAM enables fast and stable on-line learning of new input vectors [241]. This is important, as it might be necessary to integrate new data during the application of the proposed cell recognition approach. Nonetheless, an SFAM network encompasses only three layers (see Figure 5.14), which is a significant reduction in comparison to the fuzzy ARTMAP.

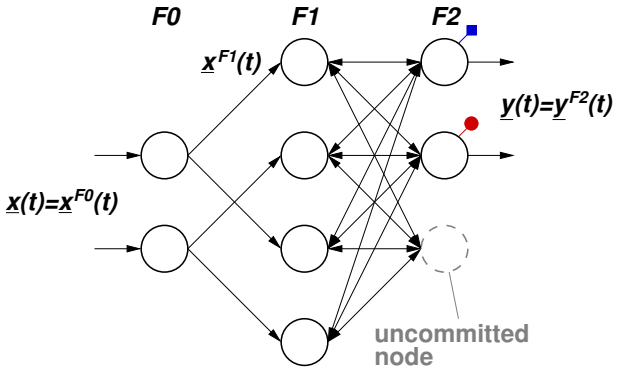


Figure 5.14: *Structure of the SFAM.* The SFAM encompasses three layers: $F0$, $F1$ and $F2$. The neurons of the $F2$ layer are associated with class labels. Furthermore, there are uncommitted nodes, which can be incorporated if new input vectors are to be learned.

The first layer $F0$ performs a suitable encoding of the input vector $\underline{x}(t)$ called complement coding. The resulting vector $\underline{x}^{F1}(t)$ constitutes the input vector of the subsequent layer $F1$. The nodes of the output layer $F2$ are linked to all nodes of the $F1$ layer. The corresponding weights $\underline{w}_i^{F2}(t)$ define hyper-rectangular subspaces of the input space – the categories. In addition, each $F2$ neuron is associated with a class label. As the SFAM is an incremental network, there are neurons, which are not in use but required for an extension of the network – the uncommitted nodes.

After a new training sample $\underline{x}(t)$ has been presented and complement coded, the $F2$ nodes i are activated according to Equation 5.42.

$$z_i^{F2}(t) = \frac{\left|\underline{x}^{F1}(t) \wedge \underline{w}_i^{F2}(t)\right|_1}{\alpha + \left|\underline{w}_i^{F2}(t)\right|_1} \tag{5.42}$$

The choice parameter α should be set slightly higher than zero. This enables small categories to be preferred to large ones. '$\wedge$' symbolises an element-wise minimum operation. The norm $|\cdot|_1$ for p-dimensional vectors $\underline{x}$, which is sometimes referred to as the city block norm, is defined as follows:

$$|\underline{x}|_1 = \sum_{k=1}^{p} |x_k| \tag{5.43}$$

After all $F2$ nodes have been activated, the best-matching category corresponding to the node j with the highest activation and the correct class label is selected. But it is only allowed to grow and enclose the new input vector if the vigilance criterion is fulfilled (see Equation 5.44). Thus,

the category size is limited by the vigilance parameter ρ.

$$\frac{\left|\underline{x}^{F1}(t) \wedge \underline{w}_j^{F2}(t)\right|_1}{\left|\underline{x}^{F1}(t)\right|_1} \geq \rho \tag{5.44}$$

Assuming a neuron was not able to fulfil the vigilance criterion, its activation is reset. Then a new best-matching node j is chosen and, in the case that the vigilance criterion is met for j, its weights are adapted according to Equation 5.45. If no suitable node is available, an uncommitted node is selected and associated with the input vector's class label.

$$\underline{w}_j^{F2}(t+1) = \eta\big(\underline{x}^{F1}(t) \wedge \underline{w}_j^{F2}(t)\big) + (1-\eta)\underline{w}_j^{F2}(t) \tag{5.45}$$

Here, a new parameter, the learning rate η, appears. Within the scope of this thesis, the learning rate is usually set to 1, in order to run the SFAM in the *fast-learning mode*. In this mode, a new input vector is enclosed by a category after one learning step. Other values for η would increase the number of training cycles significantly.

In order to classify new patterns, the activation of all $F2$ nodes is computed and the output of the best-matching node $y_j^{F2}(t)$ is set to its class label. The other $m-1$ outputs are set to -1. Then, a classification result $c(t)$ can be determined:

$$c(t) = \max_{i=1,\ldots,m} y_i^{F2}(t). \tag{5.46}$$

For the introduced cell recognition task, a measure specifying whether an input vector is known or unknown is required to reject non-cell segments. Furthermore, it would be beneficial to have information on the degree of knowledge; e.g. in order to merge the results from different types of microscope images. The activation $z_i^{F2}(t)$ cannot fulfil this task, since it varies depending on a category's size (cf. Equation 5.42). Therefore, I employed $\tilde{z}_i^{F2}(t)$ as an alternative measure for the activation, if an input vector is to be classified (cf. Equation 5.47).[4] Here, $\tilde{z}_i^{F2}(t)$ corresponds to the distance from an input vector to category i.

$$\tilde{z}_i^{F2}(t) = \left|\big(\underline{x}^{F1}(t) \wedge \underline{w}_i^{F2}(t)\big) - \underline{w}_i^{F2}(t)\right|_1 \tag{5.47}$$

The minimum value $\tilde{z}_{\min}^{F2}(t)$ of $\tilde{z}_i^{F2}(t)$ over all nodes i indicates the degree of knowledge about an input vector. Assuming $\tilde{z}_{\min}^{F2}(t)$=0, the input vector is completely known. Higher values correspond to less knowledge. However, an input vector which is close to a category is likely to be representable by it. Therefore, I have introduced a threshold τ, which denotes the maximum distance an input vector is considered as being known.

If an input vector is to be classified, the outputs $y_i^{F2}(t)$ of all F2 nodes i are set to -1. Afterwards, the input vector is complement coded and all F2 neurons are activated according to Equation 5.47 and the best-matching neuron j is determined. In the case that $\tilde{z}_{\min}^{F2}(t)$ is smaller than τ, the output y_j^{F2} of the best-matching neuron receives the input vector's class label (see Figure 5.15). Otherwise, y_j^{F2} remains unchanged. Then, $c(t)$ yields a class label for known input vectors and -1 otherwise (see Equation 5.46).

As it cannot be guaranteed that the city block norm, which is employed by the SFAM, fits to the classification task, an alternative simplified ARTMAP network was utilised for comparison.

[4] During training, the original activation (see Equation 5.42) is utilised further on.

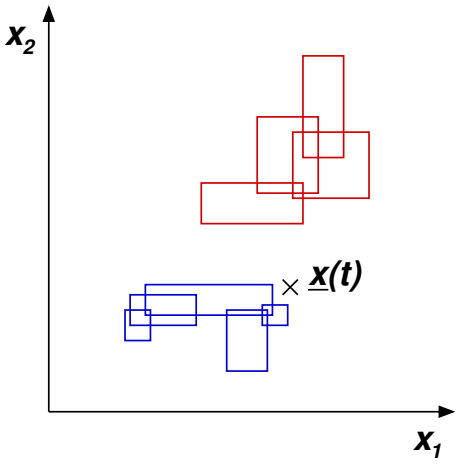


Figure 5.15: *Classification using the modified SFAM.* A new input vector $\underline{x}(t)$ is presented to an SFAM network, which performs a separation of two classes in a two-dimensional feature space. The categories belonging to each of the classes are depicted in blue and red, respectively. In principle, $\underline{x}(t)$ would be unknown to the network, as it does not lie inside any category. However, the consideration of $\tilde{z}^{F2}_{\min}(t)$ enables the input vector to be assigned to the blue class, if its distance to the next blue category is smaller than the threshold τ.

In order to obtain such a network, I simplified the structure of the hypersphere ARTMAP, which had been introduced by Georgios C. Anagnostopoulos and Michael Georgiopoulos [8], according to the SFAM. It shares the fuzzy ARTMAP's ability of fast and stable incremental learning. But instead of defining hyperrectangular categories, the weights $\underline{w}_i^{F2}(t)$ of the hypersphere ARTMAP specify hyperspheres; in particular, they comprise a category's centroid $\underline{\mu}_i$ as well as its radius R_i. This alternative encoding renders the $F0$ layer dispensable.

The simplification of the hypersphere ARTMAP network's architecture essentially boils down to a usage of the structure of the SFAM and a substitution of the equations for computing the $F2$ neurons' activations, the adaptation of the $F2$ nodes' weights and the vigilance criterion (cf. equations 5.42, 5.44 and 5.45). The new computations, which could be transferred from [8], are based on the Euclidean norm rather than the city block norm. Equation 5.48 and Equation 5.49 show the new activation and the corresponding vigilance criterion.

$$z_i^{F2}(t) = \frac{R^* - \max\left(R_i, \left|\underline{x}^{F1}(t) - \underline{\mu}_i(t)\right|_2\right)}{R^* - R_i + \alpha} \tag{5.48}$$

$$1 - \frac{\max\left(R_i, \left|\underline{x}^{F1}(t) - \underline{\mu}_i(t)\right|_2\right)}{R^*} \geq \rho \tag{5.49}$$

The parameters α and ρ are known from the SFAM. So there is only one additional quantity: the maximum category radius R^*. As the vigilance already restricts the category growth, this parameter seems to be redundant. In addition, the vigilance of the SFAM has specific properties, for instance, if it equals 0, one category is able to cover the complete input space. In order to realise equal behaviour with hypersphere ARTMAP, I set R^* to $\frac{1}{2}\sqrt{p}$, where p denotes the number of employed features. Then, there is no need to chose R^* anymore.

Similar to α and ρ, the effects of the learning rate η resemble the SFAM. Therefore, it was selected to equal 1 as well. I refer to the resulting network as the *simplified hypersphere ARTMAP* (SHAM).

Since with respect to the examined cell recognition task, unknown feature vectors are to be rejected, I had to perform an adaptation of $\tilde{z}_i^{F2}(t)$ (cf. Equation 5.47) according to the SHAM's

characteristics. The result is shown in Equation 5.50.

$$\tilde{z}_i^{F2}(t) = \max\left(\left|\underline{x}^{F1}(t) - \underline{\mu}_i(t)\right|_2 - R_i, 0\right) \tag{5.50}$$

Equation 5.50 accounts for the hyperspherical shape of the categories as well as the Euclidean norm. However, it is a measure for the distance from an input vector to a category like in the proposed SFAM approach. Therefore, a threshold τ can be applied to reject unknown patterns, whose distances to the respective closest categories are too large.

Support Vector Classification

In addition to simplified ARTMAP neural networks, I employed *support vector machines* (SVMs) [194, Chapter 7][243] to reject non-cell segments. The foundations of this class of statistical learning architectures were laid by the Russian mathematician Vladimir Naumovich Vapnik in the 1960s [194, Chapter 1][244]. He developed an alternative learning procedure for a class of simple neural networks called *perceptrons* [275, Chapter 7], which were very popular at that time. In principle, these perceptrons adapt a hyperplane in such a way as to separate distinct regions of the input space (see Figure 5.16). Vapnik expressed the parameters of this hyperplane in a form, which only depends on a set of available samples, the labels of which are known. Like in the case of the kernel PCA, this representation is named *dual representation* (see Section 5.4.3).

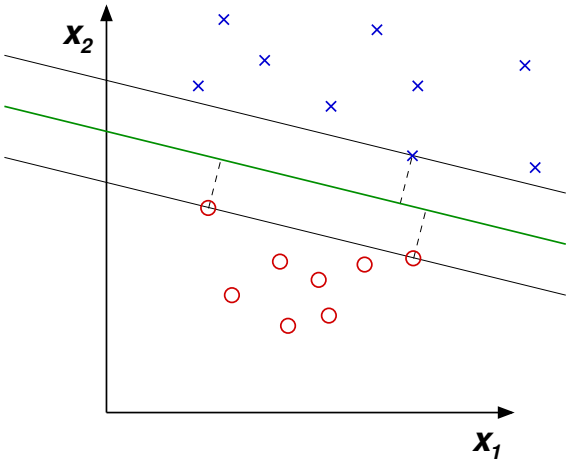


Figure 5.16: *Classification using hyperplanes.* A hyperplane depicted as a green line separates input vectors of two classes, which are shown as blue crosses and red circles, respectively. The distance of the points closest to the hyperplane, called support vectors, is referred to as margin. It has been visualised as black lines.

The decision function, which corresponds to the output $y\big(\underline{x}(t)\big)$ of such a classifier, is shown in Equation 5.51. In contrast to simplified ARTMAP networks, it can only distinguish between two classes represented by $y\big(\underline{x}(t)\big)=-1$ and $y\big(\underline{x}(t)\big)=1$, respectively. The weights $\underline{w}$ as well as the bias θ define a hyperplane according to Equation 5.52. 'sgn' denotes the sign function.

$$y\big(\underline{x}(t)\big) = \operatorname{sgn}\Big(\big\langle \underline{w}, \underline{x}(t)\big\rangle + \theta\Big) \tag{5.51}$$

$$\big\langle \underline{w}, \underline{x}(t)\big\rangle + \theta = 0 \tag{5.52}$$

Exploiting the dual representation, which has the same solution as the primal one, the decision function can be rewritten as shown in Equation 5.53. The new formulation is dependent on the n training samples $\underline{x}_i$. The label of a training pattern is symbolised by y_i. Furthermore, α_i denotes

the dual variables (Lagrange multipliers), which are a result of the alternative representation of the problem.

$$y\big(\underline{x}(t)\big) = \operatorname{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i \big\langle \underline{x}(t), \underline{x}_i \big\rangle + \theta\right) \tag{5.53}$$

The factors α_i can be determined by maximising the quantity $W\big(\underline{\alpha}\big)$ (see Equation 5.54) subject to the constraints summarised by Equation 5.55.

$$W\big(\underline{\alpha}\big) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \big\langle \underline{x}_i, \underline{x}_j \big\rangle \tag{5.54}$$

$$\forall i \in \{1, \ldots, n\} : \ \alpha_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{5.55}$$

The input vectors $\underline{x}_i$, whose α_i is non-zero, are termed *support vectors*. They constitute the points, which are closest to the separating hyperplane. Their distance to this plane is called *margin*. The other input vectors do not have any influence on the position of the hyperplane; i.e., the hyperplane is exclusively determined by the support vectors. In order to achieve a good generalisation, this margin is maximised.

The value of the bias θ can be derived using the support vectors [194, Section 7.3]. Then the decision function $y\big(\underline{x}(t)\big)$ becomes computable for unknown input vectors.

Although this kind of classifier can be used in order to discriminate data from two classes, it has two major disadvantages: Firstly, it only functions properly, if the considered points are linearly separable. Unfortunately, this is frequently not true, even for very simple tasks such as the XOR problem [275, Chapter 7].[5] Secondly, data originating from real tasks are usually noisy, which causes points from both classes to overlap. These problems are dealt with in the following.

The decision function shown in Equation 5.53 contains the computation of a dot product between two input vectors. As discussed in Section 5.4.3, such a computation can be substituted by a kernel. This application of the *kernel trick* maps all data into a higher-dimensional space, where the classification is performed. But a linear separation in the new space might be non-linear with respect to the original input space; i.e., non-linear hyperplanes can be constructed in the original input space. Depending on the task at hand, the classification might be facilitated noticeably by that.

Three important types of *kernels* are polynomial kernels, *radial basis function* (RBF) kernels and sigmoid kernels (see Equations 5.56 to 5.58) [194, Section 2.3]. Provided the degree d of the polynomial kernel is set to 1, it equals the original dot product, which is also called linear kernel [61]. Furthermore, if the RBF parameter γ equals $\frac{1}{2\sigma^2}$, the RBF kernel is called a Gaussian RBF kernel.

[5] The XOR problem regards two-dimensional binary input vectors. The class labels of these input vectors correspond to their combination using an exclusive OR operation. The input vectors that have been labelled in such a way cannot be separated by a single line, which constitutes a separating hyperplane in the considered input space.

$$\text{polynomial kernel:} \quad k\big(\underline{a},\underline{b}\big) = \big\langle \underline{a},\underline{b}\big\rangle^{d} \quad \text{, with } d \in \mathbb{N} \tag{5.56}$$

$$\text{RBF kernel:} \quad k\big(\underline{a},\underline{b}\big) = \exp\Big(-\gamma|\underline{a}-\underline{b}|_2^2\Big) \quad \text{, with } \gamma > 0 \tag{5.57}$$

$$\text{sigmoid kernel:} \quad k\big(\underline{a},\underline{b}\big) = \tanh\Big(\gamma\big\langle \underline{a},\underline{b}\big\rangle + \vartheta\Big) \quad \text{, with } \kappa < 0 \text{ and } \vartheta \in \mathbb{R} \tag{5.58}$$

But these kernels are just examples. There are numerous kernels which could be applied instead. All of them have specific properties, which need to be considered before an application.

The problem of overlapping points from both classes was addressed by introducing the so-called *slack variables* that allow for margin errors. But as very small errors are desired, the slack variables are minimised. The resulting classifier is able to tolerate a small fraction of data points, which are not separated correctly. Equation 5.59 shows the quantity $W'\big(\underline{\alpha}\big)$ that is maximised. It is subject to the constraints summarised by Equation 5.60.

$$W'\big(\underline{\alpha}\big) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k\big(\underline{x}_i,\underline{x}_j\big) \tag{5.59}$$

$$\forall i \in \{1,\ldots,n\}:\ 0 \le \alpha_i \le \frac{C}{n} \quad \text{and} \quad \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{5.60}$$

The resulting classifier constitutes the basic variant of a *support vector classifier* (C-SVC). In comparison to Equations 5.54 and 5.55, only the dot product has been substituted by a kernel and an additional constraint depending on a parameter C has been introduced. The slack variables have disappeared. But unfortunately, C, which controls the trade-off between minimising the classification error and maximising the margin, is difficult to choose [194, Section 7.5]. Therefore, Bernhard Schölkopf and his co-researchers have introduced an alternative support vector classifier – the ν-SVC [195].

While the optimisation problem is slightly altered due to this modification, the decision function remains unchanged. But the parameter C is substituted by a new parameter $\nu \in]0,1]$. Compared to C, the influence of ν on the classification is more intuitive. It defines an upper bound on the fraction of margin errors[6] and a lower bound on the fraction of utilised support vectors [195]. Through this, the choice of relevant parameters is alleviated in comparison to a C-SVC.

The actual optimisation can be realised by numerous methods, an overview of which is given in [194, Chapter 10]. The majority of these methods is applied in batch mode; i.e., the whole training set is required in order that the problem becomes solvable. But there are approaches to on-line learning as well [27][53][213]. In comparison to the batch-learning approach, the accuracy of the on-line SVCs is slightly decreased. Therefore, I decided to utilise a batch-learning technique, since I wanted to apply support vector classifiers as reference learning architecture; in particular, I employed a *sequential minimal optimisation* technique [61], which has been provided by a software called LIBSVM [30].[7] The sequential minimal optimisation constitutes a method that iteratively optimises pairs of Lagrange multipliers. It has been reported to be very fast and does not require a large amount of memory.

[6] points that are incorrectly classified or lie within the margin

[7] Within the scope of my thesis, I employed version 2.84 of LIBSVM.

Using the method detailed above, segments showing a cell can be separated from non-cell segments by means of the determined hyperplane. However, if a type of segment appears, which does not resemble any training pattern, the result is undefined. Since it is not possible do obtain samples representing all types of non-cell segments, these unknown segments are very likely to be objects other than cells. Therefore, they must be rejected. In principle, an unsupervised ART network would be capable of solving this task efficiently. However, I wanted to apply an SVM-based approach here, since the discrimination of cell and non-cell segments is realised by this kind of technique. The methods based on ART or rather ARTMAP networks are evaluated separately. With respect to SVMs, the problem of unknown patterns has been summarized under the term of *novelty detection*. It can be solved using a *single-class SVC*, which was introduced by Bernhard Schölkopf and his colleagues as well [192][194, Chapter 8].

Single-class SVCs are algorithms allowing for the discrimination of data, which lies within a small region, from all input vectors outside. This region is specified by means of training samples. In contrast to the basic SVC approach that distinguishes between two classes, no second class of data is available. Therefore, the criteria for constructing the hyperplane had to be modified. The new goal consists in constructing a hyperplane that separates the training samples from the origin with a maximum distance (see Figure 5.17). Like with the two-class SVC, slack variables were introduced. So, it is possible to neglect outliers.

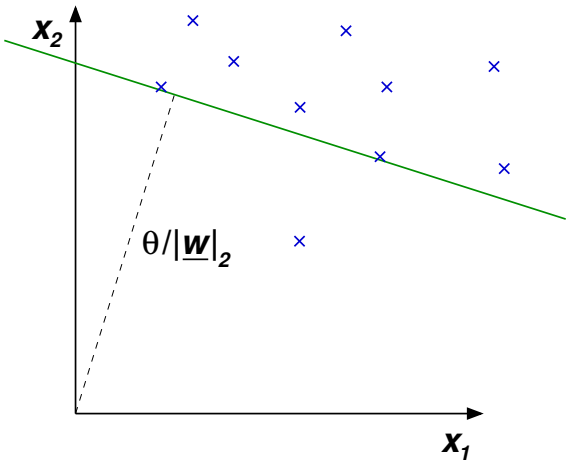


Figure 5.17: *Novelty detection using a SVC.* A hyperplane depicted as green line separates the training samples from the origin. Here, points that do not fit the volume occupied by the majority of the data can be neglected.

As Figure 5.17 shows, a hyperplane in the original input space cannot enclose the data. Therefore, the data ought to be mapped in an alternative space $\mathcal{H}$. In terms of kernels, this yields the following decision function:

$$y\big(\underline{x}(t)\big) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i k\big(\underline{x}(t), \underline{x}_i\big) + \theta\right) \tag{5.61}$$

With respect to single-class classification, in particular, the usage of RBF kernels is beneficial. An RBF kernel maps all input vectors on a hypersphere in the new space $\mathcal{H}$ that is centred at the origin. Such a distribution can be easily separated from the origin by means of a hyperplane.

Similar to a ν-SVC, a single-class SVC is controlled by a parameter $\nu \in]0, 1]$. It defines an upper bound on the fraction of outliers and a lower bound on the fraction of utilised support vectors. In the case that ν approaches zero, virtually no outliers are allowed.

On the whole, a two-stage approach was applied so as to perform a separation of cell and

non-cell segments using SVMs. Firstly, a single-class SVC decides, whether an input vector is known or not. If it is known, a ν-SVC chooses the appropriate class. Otherwise it is rejected. This means that in contrast to the ARTMAP-based approach discussed above, two classifiers need to be trained. Each of these two classifiers has its own parameters to be set. Therefore, the application of simplified ARTMAP networks, which perform both tasks at once, is more convenient. Nevertheless, the SVM-based approach is utilised as a reference method.

5.4.5 Results

Within the scope of the evaluation, the introduced feature reduction methods (see Section 5.4.3) as well as the discussed classifiers (see Section 5.4.4) were compared; i.e., each classifier (the SFAM, the SHAM and the ν-SVCs) was examined using several feature reduction techniques (CA, CA & PCA and CA & ICA) as well as the original features. Here, two variants of the ν-SVCs were considered. As detailed above, the RBF kernel is very well-suited for the single-class classification task. Hence, it was applied to this task. However, it might not fit the two-class classification problem. Therefore, a linear kernel was employed in addition to the RBF kernel in order to solve the two-class problem. The advantage of the linear kernel consists in its simplicity, which makes it very fast to compute.

The classifiers' training was performed using datasets created according to Section 5.4.2. Here, two similarity criteria for the automatic generation of segments were applied: $E_j^{\text{seg}}{=}\max\left(A_j, d_j\right)$ and $E_j^{\text{seg}}{=}d_j$. The resulting numbers of samples for each class and type[8] are shown in Table 5.2. As only the similarity measures for the automatic data generation vary, the same sets of cell masks and manually selected non-cell segments were employed.

	manually determined		automatically generated		
E_j^{seg}	cells	non-cells	cells	non-cells	total
$\max\left(A_j, d_j\right)$	759	3,119	21,967	95,536	121,381
d_j	759	3,119	56,137	80,409	140,424

Table 5.2: *Available training samples.* The set of training samples was considerably increased by means of the automatic segment generation approach. But is must be kept in mind that the automatically acquired data are not independent from the manual segmentations. Furthermore, the generation technique using A_j is more strict, which leads to a lower number of obtained cell samples.

The regarded classifiers were compared based on three criteria: the *false negative ratio* (FNR), the *false positive ratio* (FPR) and the *total accuracy* (ACC). The false negative ratio denotes the fraction of cell segments that are incorrectly classified as non-cells. In contrast, the false positive ratio indicates non-cell segments, which are regarded as cells. Both are error measures and should be minimised. As the determined cell segments are to be employed for the subcellular localisation of proteins in corresponding fluorescence micrographs, the FPR is more important than the FNR; missed cell segments do not have negative effects on the outcome of the protein localisation, whereas non-cell segments could lead to completely different results. In addition to these error measures, the total accuracy is applied so as to assess the overall correctness independent of the class; therefore, it should be high.

[8] either 'manually determined' or 'automatically generated'

Both the FNR and the FPR are intimately related to other measures reflecting the performance of classifiers. The *true positive ratio* (TPR), also called *sensitivity*, represents the fraction of cells segments that are classified as cells. It corresponds to $1-$FNR. Similarly, the *true negative ratio* (TNR), which is sometimes referred to as *specificity*, can be derived: TNR$=1-$FPR. It denotes the fraction of correctly classified non-cell segments.

The evaluation was carried out by means of ten-fold *cross-validation* [62]: The basic dataset comprising all samples obtained using the respective similarity criterion was split into ten disjoint smaller training sets, nine of which were employed for training. The remaining one constitutes the validation set. The training process was repeated for all possible validation sets and the results were averaged. Due to the amount of non-cell samples, the non-cells considered as cells have stronger effects on the accuracy in comparison to unrecognised cells. This is desirable, as incorrectly classified non-cells impair the subsequent localisation of proteins in corresponding fluorescence micrographs noticeably, whereas the influence of unrecognised cells is only limited inasmuch as they are not available.

But it had to be considered that the training samples are not independent. The automatically generated segments necessarily depict cells which have been manually extracted before. Therefore, each cell mask received a unique ID, which was transferred to the derived segments as well. The training sets were created with respect to these IDs. So, only segments depicting cells that had not been used for training were employed for validation. The classification results of the simplified ARTMAP networks are shown in Table 5.3. Although the training of the classifiers took place using manually and automatically generated samples, the evaluation is solely based on manually determined segments, as their biological relevance is higher.

		$E_j^{\text{seg}} = \max\left(A_j, d_j\right)$				$E_j^{\text{seg}} = d_j$			
classifier	feature reduction	p	FNR	FPR	ACC	p	FNR	FPR	ACC
SFAM	–	111	0.248	0.030	0.927	111	0.144	0.084	0.904
	CA	81	0.257	0.029	0.926	79	0.155	0.082	0.904
	CA & PCA	28	0.224	0.028	0.933	28	0.149	0.072	0.913
	CA & ICA	28	0.237	0.014	0.942	28	0.138	0.053	0.930
SHAM	–	111	0.336	0.047	0.896	111	0.212	0.102	0.876
	CA	81	0.294	0.046	0.906	79	0.235	0.100	0.873
	CA & PCA	28	0.292	0.039	0.911	28	0.149	0.108	0.884
	CA & ICA	28	0.377	0.013	0.916	28	0.225	0.059	0.908

Table 5.3: *Classification results of the SFAM and the SHAM.* For each combination of a classifier, a feature reduction method and a similarity criterion, the number of features p, the FNR, the FPR and the total accuracy ACC are given. The best accuracies are highlighted red. In order to enable an assessment of their differences' statistical significances, the respective confidence intervals are shown in Table C.2.

The numbers given in Table 5.3 constitute the best results obtained by the respective classifiers using different parameter settings. Since both classifiers can be completely characterised by the vigilance ρ and the threshold τ, only two parameters have to be considered. ρ was iterated in the interval $[0.95, 1.00[$ that has proven to be sufficient in preliminary experiments. The stepsize was set to 0.001. With respect to τ, values from the interval $[0, 5]$ were considered using a stepsize of 0.05. But, as τ is merely relevant for classifying data but not for training, it can be determined efficiently; only one classifier needs to be trained so as to compute the results for all values of τ.

Based on the results given in Table 5.3, it can be concluded that both types of classifiers allow for a discrimination between cell and non-cell segments. However, the highest accuracies reached by the SHAMs are significantly lower than the corresponding results of the SFAM (significance level: α=0.05, cf. Table C.2). Therefore, the SFAM is more convenient with respect to the task at hand.

Besides the type of classifier, the feature reduction method has a strong influence on the quality of the discrimination. The usage of the CA in conjunction with the ICA yielded the highest accuracies. Except for the combination of the SFAM with the criterion $E_j^{\text{seg}}=\max\left(A_j, d_j\right)$, these accuracies are even significantly higher in comparison to the results for the respective non-reduced feature set (significance level: α=0.05, cf. Table C.2). So, the reduction of the feature spaces alleviated the classification task, even though less information was available. Apparently, the new arrangement of the data fitted the characteristics of the classifiers. In particular, the simplified ARTMAP networks seem to benefit from small sets of independent features.

The pattern generation method affects the results as well. Although the highest accuracies do not differ considerably for both similarity criteria, the relation of the FNR and the FPR changes noticeably. $E_j^{\text{seg}}=\max\left(A_j, d_j\right)$ leads to FPRs which are very low and reach values down to about 1%. This would be beneficial with respect to the subsequent protein localisation, as almost all obtained segments represent real cells. On the other hand, the FNRs are considerably higher. So, $E_j^{\text{seg}}=d_j$ might be the better choice if too few cells are recognised (see Section 5.5), even though the FPRs triple and reach values up to 6%. However, 6% are still acceptable, since the final number of incorrectly classified segments is further decreased by an additional mechanism (see Section 5.5)

The experiments regarding the SVCs were conducted in an identical way using two different types of classifiers. Both employ a RBF kernel for the single-class problem. But the two-class problem is tackled using an RBF as well as a linear kernel. Since only the kernel function of the second problem varies, the kernel of the single-class classifier is usually omitted if the classifiers are referred to.

In contrast to the simplified ARTMAP networks, the classification depends on a considerably higher number of parameters: the training parameters ν_1 and ν_2 for the single-class and the two-class SVC, respectively, as well as the kernel parameters γ_1 and γ_2 for the corresponding RBF kernels. In the case that a linear kernel is utilised for the discrimination between cell and non-cell segments, γ_2 is not required. So the SVC's training depends on three to four parameters rather than one (the vigilance ρ) employed by the simplified ARTMAP networks.

By virtue of this increased number, the regarded intervals had to be strictly limited. ν_1 received the value of 0.001, which means that at least 99.9% of the training data is lying within the subspace of known samples. So, virtually no training samples are classified as unknown, which resembles the method for choosing the neural networks' parameter τ, since it maximises the total accuracy as well. In contrast to ν_1, the values of ν_2 were iterated in the interval $]0, 0.1]$. The upper limit was chosen in such a way as to avoid margin errors larger than 10% of the training set size. Here, the accuracies achieved by the simplified ARTMAP networks served as a basis for the estimation of likely results. If such values had not been available, a more comprehensive investigation could not have been avoided. The parameters γ_1 and γ_2 are the most difficult to set, as they are not associated with an intuitive meaning and exclusively depend on the data. Thus, they were chosen from the set $\{0.0001, 0.001, 0.01, 0.1, 1.0\}$, which covers several orders of magnitude. The corresponding cross-validation results are summarised by Table 5.4.

classifier	feature reduction	$E_j^{\text{seg}} = \max(A_j, d_j)$ p	FNR	FPR	ACC	$E_j^{\text{seg}} = d_j$ p	FNR	FPR	ACC
ν-SVC, linear	–	111	0.144	0.037	0.942	111	0.071	0.101	0.905
	CA	81	0.097	0.067	0.927	79	0.087	0.083	0.916
	CA & PCA	28	0.202	0.037	0.931	28	0.076	0.100	0.905
	CA & ICA	28	0.204	0.042	0.926	28	0.116	0.099	0.898
ν-SVC, RBF	–	111	0.150	0.036	0.942	111	0.091	0.079	0.919
	CA	81	0.137	0.050	0.933	79	0.117	0.085	0.909
	CA & PCA	28	0.198	0.038	0.930	28	0.088	0.105	0.899
	CA & ICA	28	0.202	0.037	0.931	28	0.084	0.098	0.905

Table 5.4: *Classification results of the SVCs.* For each combination of a support vector classifier, a feature reduction method and a similarity criterion, the number of features p, the FNR, the FPR and the total accuracy ACC are given. The best accuracies are highlighted red and the respective confidence intervals are shown in Table C.3.

From Table 5.4 it can be concluded that both types of kernels enable an excellent discrimination of the two segment classes. There is no statistically significant difference in the accuracy of the respective classifiers (significance level: α=0.05, cf. Table C.3). The best results are comparable to the SFAM, although individual accuracies, e.g. for the combination of the CA and the ICA using E_j^{seg}=d_j, differ significantly.

In contrast to the simplified ARTMAP networks, the feature reduction led to a decrease of the accuracies. So, with one exception (linear ν-SVC using E_j^{seg}=d_j), no advantage could be taken from using a reduced feature set that is faster to compute than the complete one. For some classifiers, this decline is even statistically significant (significance level: α=0.05, cf. Table C.3), which might be a drawback if a high-throughput application is the goal. This behaviour could partially result from the fact that the feature vectors are implicitly transferred into a new feature space. So, the properties of the original feature space are neglected. However, the results regarding the system using the linear kernel indicate that the SVCs are not able to take advantage of these properties, even if the feature space is not modified.

The influence of the similarity criterion used for the automatic generation of patterns seems to have effects known from the simplified ARTMAP networks: E_j^{seg}=$\max(A_j, d_j)$ enables lower FPRs than E_j^{seg}=d_j. But since the overall accuracies are similar, E_j^{seg}=d_j leads to lower FNRs as well. This shows that the pattern generation technique is able to control the classification results independent of the classifier used. However, in comparison to the SFAM and the SHAM, the FPRs seem to be higher and the FNRs appear lower. But in principle, a similar effect could be realised by modifying the similarity criterion.

On the whole, the SFAM and the SVCs proved their applicability to the task of discriminating between cell and non-cell segments. The SHAM performed significantly worse; therefore, it should not be utilised. Moreover, the SFAM has several advantages in comparison to the SVCs: First of all, it enables fast incremental learning during its application. Secondly, it performs both classification tasks at once, which leads to a noticeably reduced number of parameters to be set. Finally, it often performs better using the reduced feature set. So the computational load can be slightly diminished. Whilst the SFAM, on average, requires 0.158s to classify a segment using an AMD Athlon 64 processor operating at 2GHz, the SVCs take 0.176s and 0.156s using an RBF kernel and a linear kernel, respectively. These values refer to the respective best classifiers using

$E_j^{\text{seg}}=d_j$ and include the time required for computing the features. Here, the SVC employing the linear kernel utilised the feature set reduced by CA. Therefore, the computation time is similar to the SFAM. But the best results were achieved by the SHAM that performs the classification in 0.146s per segment.

5.5 Evaluation of the Complete System

All experiments detailed above concern special aspects of the cell recognition method, in particular the localisation, segmentation and classification. They were performed using cross-validation. However, the knowledge gained in the respective preceding steps was applied; for instance, the classification takes advantage of the segmentation procedure so as to acquire additional training data. Therefore, there is necessarily a bias caused by the employed datasets. In order to circumvent problems resulting from this bias and to enable a more objective assessment of the complete system, a final evaluation using an independent test set of 19 bright-field images was performed. From these images 302 cell masks had been extracted by a biological expert. These cell masks served as reference.

But before cells could be recognised and associated with cell masks, another problem had to be solved. Depending on the computed cell markers, an individual cell can be segmented by multiple snakes. Therefore, it might be recognised several times, which must be avoided. In principle, the best segment could be chosen in this case. But in order to achieve this goal, a value reflecting the confidence of each classification result is required. Regarding the simplified ARTMAP networks, the minimum $\tilde{z}_{\min}^{F2}(t)$ of the modified activation $\tilde{z}_i^{F2}(t)$ over all nodes i is applicable here. It indicates the distance from an input vector in the feature space to the next category. Unfortunately, the SVCs do not provide a measure that is equally intuitive. Thus, the best segment is chosen arbitrarily there, for example, by selecting the one with the lowest index.

The confidence value provided by the simplified ARTMAP networks, enables not only a selection of the most appropriate snakes, it helps to reject incorrectly classified non-cells as well. Since the best-fitting snake is chosen, segments, which comprise multiple cells or parts thereof, are usually discarded. Thus higher FPRs can be accepted, which enables the application of classifiers trained using $E_j^{\text{seg}}=d_j$. This is a major advantage in comparison to the SVCs, in particular, since the SVCs exhibit higher FPRs (cf. Section 5.4.5).

Table 5.5 gives the fractions of recognised cells as well as the corresponding mean values $\bar{d}$ of the segmentation error d_j for all four classifiers and both methods used for automatically generating additional patterns. Regarding each classifier, the feature reduction approach which resulted in the highest accuracy was chosen.[9]

Table 5.5 reveals that the similarity criterion $E_j^{\text{seg}}=\max\left(A_j, d_j\right)$ results in insufficient recognition rates regarding the task at hand. Here, a rate of about 75% is considered satisfactory. Higher recognition rates would be beneficial, but it is not necessary to recognise all visible cells. However, rates of 50% are extremely low. So the corresponding systems should only be applied if the false positive ratios of the alternative classifiers were to high.

The recognition rates of the systems using $E_j^{\text{seg}}=d_j$ are significantly higher, but vary considerably. Since the corresponding total accuracies are similar, this is more likely to be caused by the differing false negative ratios, which explains the low recognition rates regarding

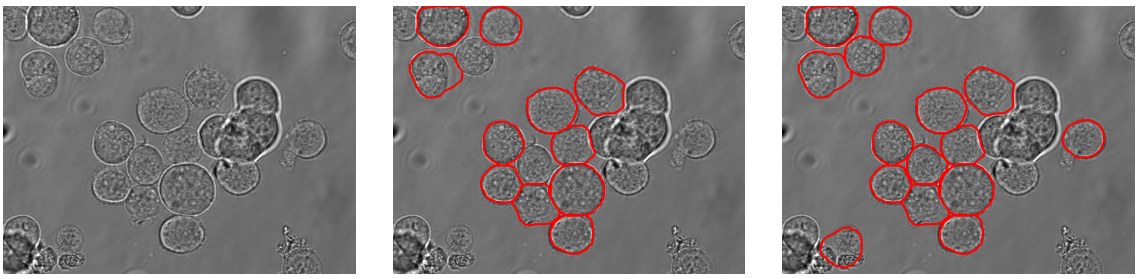
[9] The corresponding accuracies are highlighted red in Table 5.3 and Table 5.4.

	$E_j^{\text{seg}} = \max\left(A_j, d_j\right)$			$E_j^{\text{seg}} = d_j$		
classifier	rec. rate	confidence interval	$\bar{d}$	rec. rate	confidence interval	$\bar{d}$
SFAM	50.0%	[44.4%,55.6%]	0.092	81.5%	[77.1%,85.9%]	0.108
SHAM	40.1%	[34.6%,45.6%]	0.098	76.8%	[72.0%,81.6%]	0.110
ν-SVC, linear	58.3%	[52.7%,63.9%]	0.082	89.4%	[85.9%,92.9%]	0.110
ν-SVC, RBF	56.3%	[50.7%,61.9%]	0.080	90.1%	[86.7%,93.5%]	0.108

Table 5.5: *Evaluation using an independent test set.* The quality of the proposed cell recognition is assessed based on the fraction of recognised cells (rec. rate) as well as their mean segmentation errors $\bar{d}$. For determining the confidence intervals of the recognition rates, a confidence level of 0.95 was chosen (cf. Appendix C.3).

$E_j^{\text{seg}}=\max\left(A_j, d_j\right)$ as well. But the recognition rates do not reflect the quality of the cell recognition approach completely. For example, the recognition rate of the SFAM is significantly lower than the one of the SVC using an RBF kernel. But as the FPR of this SVC is higher, it can be concluded that here the cell recognition returns a higher number of incorrectly recognised non-cell segments as well. This problem is intensified, since no confidence value is provided by the SVC, which could enable a rejection of inappropriate segments. Furthermore, the FNR of the SFAM can be diminished by employing a modified similarity criterion, such as $d_j<0.11$ rather than $d_j<0.10$, for the generation of additional cell samples. But nonetheless, Table 5.5 shows that recognition rates of 90% are realisable using the suggested approach.

In order to illustrate the results, Figure 5.18 shows the cells recognised in one of the images from the test set. Here, both similarity criteria were employed.



(a) (b) (c)

Figure 5.18: *Cell recognition using an image from the test set.* A bright-field image (a) was processed using the localisation, segmentation and classification procedures discussed above. The images (b) and (c) depict the contours of those segments, which were selected as cells by means of an SFAM classifier using $E_j^{\text{seg}}=\max\left(A_j, d_j\right)$ and $E_j^{\text{seg}}=d_j$, respectively, for training. The feature reduction was realised by the combined approach employing the CA and the ICA.

The time required to recognise cells varied from 13.6s[10] to 15.1s[11] per image (1344×1024 pixels), on average. These measurements were performed using an AMD Athlon 64 processor (2GHz, 32-bit mode). Here, it should be considered that all optimisations detailed in the previous sections are based on some kind of error measure; the computation time was neglected. However, it can be decreased by relatively easy modifications. Firstly, the maximum number of iterations of the snake algorithm could be reduced (cf. Appendix C.1). Secondly, segments, which are not able to represent cells, e.g. due to their size, can be rejected before actually classifying them. Thirdly,

[10] SHAM, CA & ICA

[11] ν-SVC, RBF kernel, no feature reduction

all computations might be performed using a down-scaled image. However, these modifications require a new determination of the relevant parameters. Otherwise, suboptimal recognition rates would be the result. As the computation of segments is independent of each other, an alternative possibility to increase the algorithm's speed consists in parallelisation. Due to the current spreading of multi-core processors, this is an important property.

5.6 Adaptation to *Drosophila* Cells

Since the beginning of the 20th century, the fruit fly *Drosophila* has been an object of intense study [120]. With their investigation of *Drosophila melanogaster*, which is commonly used in laboratories all around the world, Thomas Hunt Morgan and his co-researchers laid the foundations of modern genetics (cf. Section 3.3.1). But the interest in *Drosophila* has not waned. Besides analysing fruit flies in order to examine the development of organisms [171], specialised cell lines are frequently utilised, for example, for studies concerning host–pathogen interactions [3]. Like Sf9 cells, such cell lines have been employed for expressing foreign proteins [85][224].

In 2000, the complete nucleotide sequence of *Drosophila melanogaster* was determined [1]. This knowledge about the genome constitutes a big advantage in comparison to *Spodoptera frugiperda* cells, since it alleviates the construction of fusion proteins and by that their subcellular localisation. Furthermore, predictions about possible protein locations can be made based on the corresponding genes' nucleotide sequences (cf. Section 3.3.6). Therefore, the recognition of a *Drosophila* cell line termed *S2R+* [269] has been investigated besides Sf9 cells.

Unfortunately, S2R+ cells are noticeably more difficult to segment than Sf9 cells, which can be separated from their surrounding by a clearly visible cell membrane (see Figure 5.19).

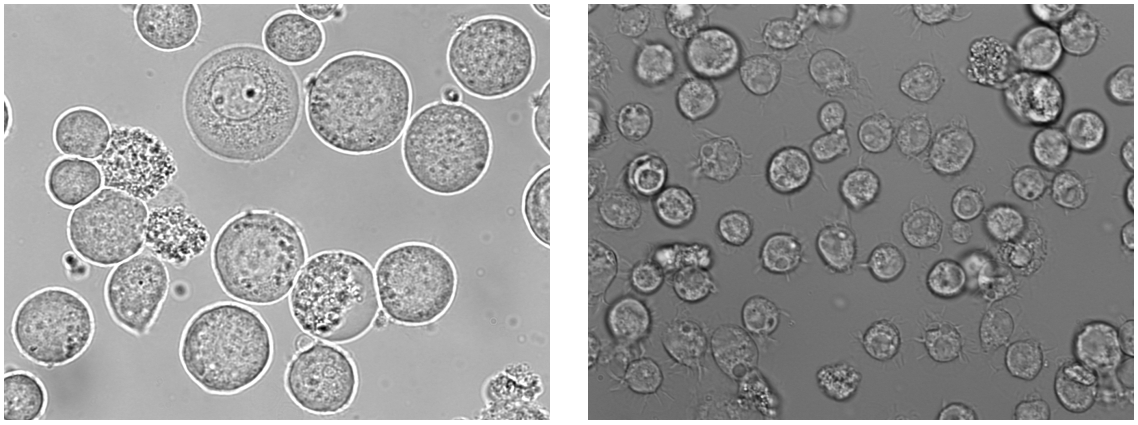


Figure 5.19: *Comparison of Sf9 (a) and S2R+ (b) cells using 60× magnification and bright-field microscopy.* The S2R+ cells are smaller than the Sf9 cells and exhibit a greater variety of shapes. Furthermore, the cell membranes are less distinct.

The cell membranes of the S2R+ cells are less distinct and the shapes are considerably more irregular. Therefore, the application of the original cell recognition approach, which was introduced in sections 5.3, 5.4 and 5.5, to bright-field images of S2R+ cells resulted in significantly worse segmentations in comparison to Sf9 cells (cf. Table 5.6). Consequently, several crucial adaptations had to be performed. Firstly, the segmentation is computed in two corresponding images: a bright-field image and a DIC image (see Figure 5.20). This does not lead to any additional constraints

as both bright-field and DIC microscopy can be applied in conjunction with fluorescence microscopy without causing any negative effects (see Section 5.2.1). But each of these images reveals details that are not visible in the other. DIC microscopy increases the overall contrast but results in discontinuous intensity levels of cell membranes, as local gradients are displayed. On the other hand, the cells are less visible in bright-field images, but pixels showing cell membranes exhibit a lower variance.

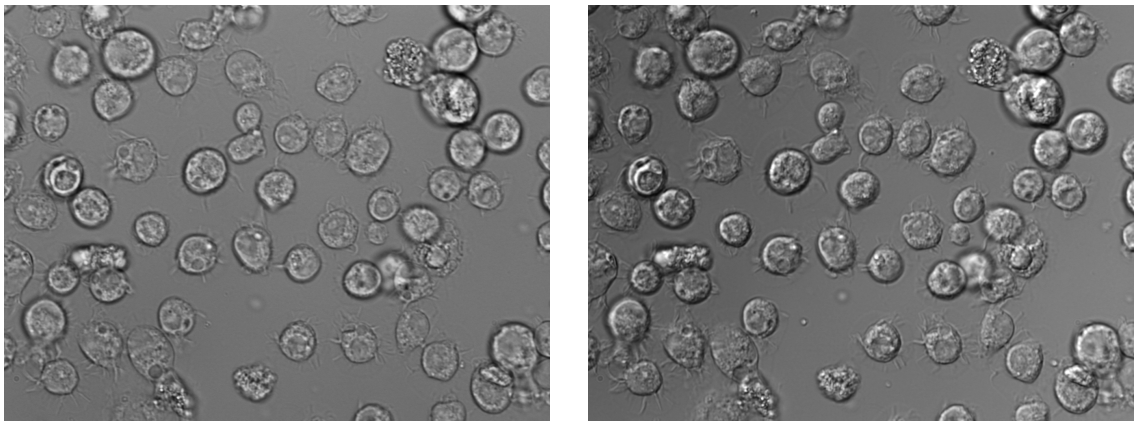


Figure 5.20: *Comparison of a bright-field (a) and a DIC image (b) depicting S2R+ cells.* The appearance of the considered cells varies depending on the employed microscopy technique. The contrast of cells that are barely visible in bright-field images is improved using differential interference contrast. But unfortunately, DIC increases the contrast of intracellular structures as well.

Due to the new cell shape and the application of two microscope images, several adaptations with respect to the segmentation and the classification were required [228]. These modifications are addressed in Section 5.6.1 and Section 5.6.2, respectively. The localisation, i.e. the determination of cell markers, was transferred unchanged. In principle, the bright-field image and the DIC image are processed separately. Alternatively, both micrographs could be regarded as a single multi-channel image and analysed in parallel. But this would necessitate comprehensive changes of the applied methods, for instance of the morphological operators and of the active contours. Therefore, the independent processing was favoured. Eventually, the results of the classification step are combined so as to yield the final recognition. The complete cell recognition approach is evaluated in Section 5.6.3.

5.6.1 Adapted Segmentation Procedure

Similar to Sf9 cells, the S2R+ cells are segmented based on cell markers that are individually computed for the considered images. In order to account for the appearance of S2R+ cells, the energy functional of the snake approach was modified. The new parameter κ enables a non-linear modification of the distance energy, for instance, in order to account for high intracellular gradient magnitudes (see Equation 5.62). If κ is set to one, this method equals the original segmentation technique for Sf9 cells (cf. Equation 5.17).

$$\begin{aligned} E^{*}_{\text{snake}}\big(c(s)\big) &= \int_0^1 \Big[\alpha E_{\text{cont}}\big(c(s)\big) + \beta E_{\text{curv}}\big(c(s)\big) \\ &\quad +\gamma\Big(E_{\text{dist}}\big(c(s)\big)\Big) E_{\text{ao}}\big(c(s)\big) \\ &\quad +\delta\Big(E_{\text{dist}}\big(c(s)\big)\Big) E_{\text{dist}}\big(c(s)\big)^{\kappa}\Big] ds \end{aligned} \tag{5.62}$$

Additionally, the computation of the energy weight $\gamma\big(E_{\text{dist}}\big)$ had to be modified accordingly (cf. Equation 5.63). In contrast, no change of $\delta\big(E_{\text{dist}}\big)$ was necessary, as its calculation is performed depending on $\gamma\big(E_{\text{dist}}\big)$ (see Equation 5.64).

$$\gamma\Big(E_{\text{dist}}\big(c(s)\big)\Big) = \gamma_0 \cdot \frac{\Delta_{\max} - E_{\text{dist}}\big(c(s)\big)^{\kappa}}{\Delta_{\max}} \tag{5.63}$$

$$\delta\Big(E_{\text{dist}}\big(c(s)\big)\Big) = \delta_0 + \gamma_0 - \gamma\Big(E_{\text{dist}}\big(c(s)\big)\Big) \tag{5.64}$$

Although the principal computations are very similar to the method which performs a recognition of Sf9 cells, the relevant parameters have to be chosen in a different way, in order to enable an optimal exploitation of both images. Furthermore, the parallel segmentation of two microscope images as well as the incorporation of the exponent κ in the energy functional (see Equation 5.62) increase the number of parameters to be set. As a result, an exhaustive search is no longer possible. Therefore, I resorted to a genetic algorithm to reduce the computational load. *Genetic algorithms* employ evolutionary mechanisms such as selection, mutation and recombination, also referred to as crossover, to a set of possible solutions [59, Chapter 9][66, Chapter 3] (see Section 3.3.1 as well). The characteristics of these solution candidates, referred to as individuals, are encoded by their genome. Concerning the task at hand, the genome $\underline{g}_i$ of an individual i consists of the relevant parameters for the segmentation of both types of images (see Equation 5.65). All these parameters are scaled to the interval $[0, 1]$. So, the evolutionary operators modify them in a similar way.

$$\begin{aligned} \underline{g}_i = \Big(&\beta^{\text{bf}}, \gamma_0^{\text{bf}}, \delta_0^{\text{bf}}, \Delta_{\max}^{\text{bf}}, \kappa^{\text{bf}}, \tau_{\text{m}}^{\text{bf}}, d_{\text{m}}^{\text{bf}}, l^{\text{bf}}, \\ &\beta^{\text{dic}}, \gamma_0^{\text{dic}}, \delta_0^{\text{dic}}, \Delta_{\max}^{\text{dic}}, \kappa^{\text{dic}}, \tau_{\text{m}}^{\text{dic}}, d_{\text{m}}^{\text{dic}}, l^{\text{dic}}\Big) \end{aligned} \tag{5.65}$$

In order to enable the application of a genetic algorithm, the quality of the segmentation must be assessed first. Here, the shape-dependent error measure d_j, which was introduced in Section 5.3.7, is utilised. Since each cell is visible in both of the considered images, one segment has to be chosen for the evaluation. Furthermore, in order to recognise a cell, its correct segmentation in one image is sufficient. Hence, I decided to select the minimum error (cf. Equation 5.66).

$$d_j^{\text{both}} = \min\left(d_j^{\text{bf}}, d_j^{\text{dic}}\right) \tag{5.66}$$

Based on d_j^{both}, the evaluation of an individual i becomes possible. Each individual is associated with a fitness value $f(\underline{g}_i)$ computed according to Equation 5.67. Here, n denotes the number of cell masks and n_{c} the number of actually localised cells.

$$f(\underline{g}_i) = 1 - \frac{1}{n_{\text{c}}} \sum_{j=1}^{n_{\text{c}}} d_j^{\text{both}} - c_{\text{m}}\left(1 - \frac{n_{\text{c}}}{n}\right) \tag{5.67}$$

The fitness function penalizes high segmentation errors and, weighted by the constant c_{m}, the number of undetected cells, as these cells cannot contribute to the segmentation error. Therefore, the genetic algorithm tends to select individuals performing an accurate segmentation of a large number of cells.

In order to cope with slight differences in the fitness values of the population, rank-based selection[12] is applied [59, Chapter 8]. Furthermore, arithmetic crossover[13] and mutation[14] for continuous-valued genes are utilised [59, Chapter 9], as the genome consists of a sequence of numerical values representing specific parameters of the approach. The crossover probability as well as the parameters controlling the rate of mutation had been determined within the scope of preliminary experiments. The final parameter settings for the segmentation approach were obtained from the best-fitting individual that occurred.[15]

The evaluation of the proposed technique was performed by comparing the mean, the standard deviation and the median of the segmentation error d_j^{both} as well as the numbers of determined cell markers and localised cell masks with the results of the original segmentation method. This comparison took place using 20 image pairs (1344×1024 pixels). From these images, 489 cells had been extracted manually. As the original technique is only able to process single images, independent experiments using either bright-field images or DIC images were conducted. In addition, investigations of the segmentation using an alternative value for κ were carried out.

All examinations were performed by means of cross-validation. The parameters for the original method were determined for 19 images and tested on the remaining one. In contrast, the genetic algorithm was evaluated using five sets of four images each: four training sets and one test set. So the computational effort for the optimisation decreased. The final values, which are shown in Table 5.6, were then obtained by averaging over subsequent trials using different test and training images.

method	n_{m}	n_{c}	μ	σ	m
original segmentation, bright-field	1660	467	0.266	0.156	0.222
original segmentation, bright-field, κ=2	1660	467	0.214	0.132	0.169
original segmentation, DIC	1190	443	0.250	0.149	0.214
original segmentation, DIC, κ=2	1190	443	0.215	0.133	0.185
genetic algorithm	1906	469	0.153	0.101	0.119

Table 5.6: *Number of determined markers n_{m}, number of localised cells n_{c} (max.: 489) as well as mean μ, standard deviation σ and median m of the segmentation error.* The genetic algorithm clearly outperforms the original approach, which uses either bright-field or DIC images. Furthermore, the new parameter κ has a strong positive influence on the segmentation.

Intriguingly, the number of cell markers resulting from the genetic algorithm is much lower than the sum of the markers extracted by the original approach in bright-field and DIC images. As a result, the total computational effort for processing these images in parallel is only moderately higher than for the usage of either image type.

[12] realised by means of non-deterministic linear sampling [59, Chapter 8]

[13] crossover probability: 0.1

[14] changes sampled from $\mathcal{N}\left(0, 0.025^2\right)$

[15] An individual is considered as best-fitting if it achieves the minimum segmentation error and localises at least 95% of all cell masks.

The segmentation results of the original approach are not satisfying; either in bright-field or in DIC images. By comparison, mean errors of about 0.11 were observed using Sf9 cells (see Tables 5.1 and C.1). Values of κ, which are higher than 1.0, entail a decreased importance of intracellular structures during the segmentation. So, the segmentation errors of the original approach could be reduced with respect to both types of images using κ=2. However, only the genetic algorithm enabled results lying in the range known from Sf9 cells. The median, in particular, indicates that the majority of these cells is segmented properly.

5.6.2 Adapted Classification Procedure

After the segmentation, the resulting segments in each image – bright-field and DIC – need to be analysed. Like with Sf9 cells, some segments represent real cells, whereas others do not. Therefore, the segments are also classified as cells and non-cells, respectively. The SFAM has proven advantageous for this purpose (cf. Section 5.4.5 and Section 5.5). It enables fast and incremental learning, an efficient rejection of surplus snakes as well as a recognition of unknown patterns, which is required, as it is not possible to acquire representative training samples for every kind of non-cell which might occur. Therefore, the SFAM is employed here as well. However, the utilised feature set had to be modified due to the higher segmentation errors caused by S2R+ cells.

The original feature set, which was tailored to the recognition of Sf9 cells, encompasses three shape features and numerous histogram-based features like statistical moments and quantiles. They are computed for several segment-specific image regions: the whole segment, its core and two overlapping tubes centred around the segment's contour with radii of 5% and 10% of the mean cell diameter, respectively (see Section 5.4.1 and Appendix B.1). In order to enable the usage of S2R+ cells, these image regions had to be modified (see Figure 5.21).

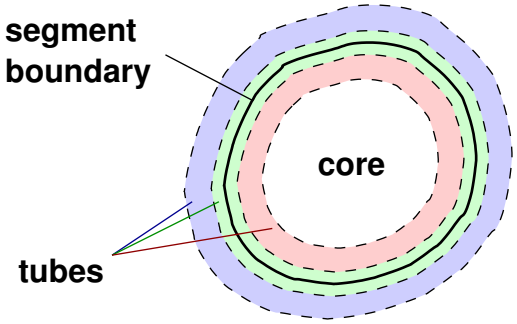


Figure 5.21: *Considered image regions.* The features are computed within five segment-specific image regions: the segment itself, three tubes around its boundary and its core.

The segment itself and its core are considered further on. But here, three non-overlapping tubes are employed, each having a radius of 5% of the mean cell diameter. So, the tubes cover a total radius of 15% around a segment's boundary, which accounts for the increased errors resulting from the segmentation of S2R+ cells (cf. Section 5.6.1 and Section 5.3.7).

Because of the higher number of considered image regions, 144 instead of 111 basic features are computed (see Appendix B.2). By virtue of the classification results obtained using Sf9 cells, these feature sets were reduced by means of the approach using a combination of the CA and the ICA, which had performed best in conjunction with the SFAM (see Section 5.4.5).

Since the manual segmentation of cells, which is required to provide training data, is very laborious, the number of training samples was increased as suggested in Section 5.4.2 and [233]. Several images were segmented using different values for the snakes' energy weights. Provided a resulting

segment showed a high similarity to the corresponding manually extracted cell ($E_j^{\text{seg}}<0.1$), it was considered as cell as well. In addition, very dissimilar segments ($E_j^{\text{seg}}\geq 0.33$) were regarded as non-cells. In contrast to Sf9 cells, here only the shape-dependent similarity measure $E_j^{\text{seg}}=d_j$ was utilised (see Section 5.3.7). The alternative measure $E_j^{\text{seg}}=\max\left(A_j, d_j\right)$ was omitted, as it appeared to be too strict for S2R+ cells, which lead to higher segmentation errors. Since the criterion $d_j<0.1$, which is responsible for the acceptance of cell segments, might have also been too strict regarding S2R+ cells, $d_j<0.125$ was considered as an alternative. Besides the automatic generation of non-cell segments, examples of non-cells were manually selected in segmented images as well. The types and numbers of applied training samples are introduced by Table 5.7.

		manually determined		automatically generated		
criterion	image type	cells	non-cells	cells	non-cells	total
$d_j < 0.1$	bright-field	489	961	5,657	12,264	19,371
	DIC	489	1,144	6,415	18,409	26,457
$d_j < 0.125$	bright-field	489	961	8,205	12,264	21,919
	DIC	489	1,144	9,401	18,409	29,443

Table 5.7: *Available training samples.* Like with Sf9 cells, the set of training samples was considerably increased by means of the automatic segment generation approach. The change of the employed similarity criterion from $d_j<0.1$ to $d_j<0.125$ caused a rise of the number of generated cell segments, whereas the other numbers remain untouched.

The evaluation of the modified classification approach was performed using 489 images of single cells (cf. Section 5.6.1). In addition, non-cell segments that had been manually selected in bright-field as well as DIC images were utilised. Since the resulting number of training samples was too small, about 40,000 additional segments depicting both cells and other image regions were generated automatically. Here, the modification of the similarity criterion from $d_j<0.1$ to $d_j<0.125$ resulted in an increase of the number of cell segments, which were generated.

Table 5.8 contrasts the classification results for the old Sf9-specific feature set and the set adapted to S2R+ cells. The numbers of basic features amounts to 111 and 144, respectively. But they were reduced by means of the CA and the ICA. The size p of the resulting feature sets is given in conjunction with the false negative ratio, the false positive ratio and the total accuracy achieved by the corresponding classifier. As with the Sf9-specific recognition technique, the evaluation is solely based on manually determined segments due to their higher biological relevance. Therefore, all results were obtained by means of ten-fold cross-validation with respect to the manually extracted segments insofar as each test set encompassed only cells which had not been used for training (cf. Section 5.4.5).

In contrast to its application to Sf9 cells where accuracies up to 93% were reported using the criterion $d_j < 0.1$ (see Section 5.4.5), the approach using $d_j<0.1$ and the Sf9-specific feature set achieved only moderate results. This is likely to originate from the segmentation procedure, whose errors are increased due to the more complex appearances of the considered S2R+ cells.

The modified feature set seems to compensate for this effect to a certain degree, although the difference is not statistically significant (cf. Table C.4). However, as I apply two classifiers in parallel, this does not constitute a drawback; it rather results in an increased importance of the FPRs which equal the resulting FPR when added together in the case of both microscopy techniques being combined. Even using this pessimistic estimation, the FPR does not reach values higher than 10% for $d_j<0.1$. More realistically, smaller false positive ratios can be expected, as only a

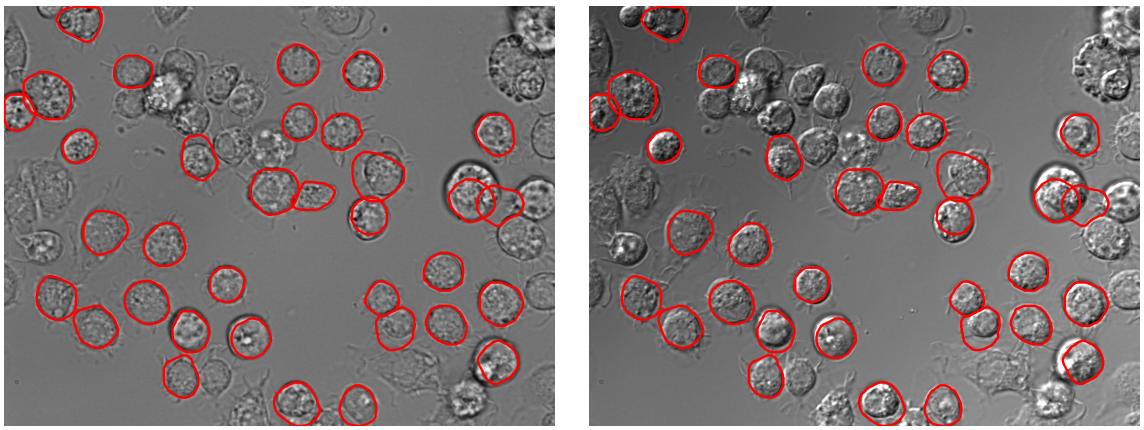
feature set	image type	similarity criterion	p	FNR	FPR	ACC
Sf9-specific	bright-field	$d_j < 0.1$	24	0.325	0.066	0.850
		$d_j < 0.125$	24	0.278	0.094	0.844
	DIC	$d_j < 0.1$	28	0.299	0.059	0.870
		$d_j < 0.125$	28	0.254	0.073	0.873
S2R+-specific	bright-field	$d_j < 0.1$	34	0.247	0.058	0.878
		$d_j < 0.125$	34	0.227	0.091	0.863
	DIC	$d_j < 0.1$	37	0.282	0.032	0.893
		$d_j < 0.125$	37	0.217	0.064	0.890

Table 5.8: *Classification results.* Number of employed features p, false negative ratio FNR, false positive ratio FPR and total accuracy ACC with respect to the manually extracted cells segments based on the Sf9-specific and the S2R+-specific feature set. In order to enable an assessment of the statistical significances of the accuracies' differences, the respective confidence intervals are shown in Table C.4.

subset of the observable objects is found in each image. So, this method can be employed for cell recognition in the context of protein localisation. Regarding $d_j<0.125$ the FPRs rise, whereas the total accuracies remain roughly equal. This results in decreased FNRs. On the whole, the technique based on $d_j<0.125$ could be applied if too few cells were recognised using $d_j<0.1$. But by virtue of the lower FPRs, $d_j<0.1$ is more convenient for the task at hand.

5.6.3 Evaluation of the Modified Approach

After the classification, the segments obtained from both images need to be combined in order to ensure that one cell is represented by not more than one segment. In principle, this problem equals the one occurring if a cell receives multiple markers and is segmented by several snakes (see Section 5.5). It is solved by associating segments, which were classified as cells and cover similar image regions. Then, the segment yielding the lowest distance $\tilde{z}^{F2}_{\min}(t)$ is selected (see Figure 5.22).



(a) (b)

Figure 5.22: *Final recognition using one of the test image pairs.* The recognised cells have been drawn as red contours into both the bright-field (a) and the DIC image (b).

76.9% of 225 reference cells, which had been manually extracted from ten independent test image pairs (1344×1024 pixels), were found by means of the S2R+-specific method with $d_j<0.1$;

i.e., a sequential application of the proposed localisation, segmentation, and classification techniques. Here, the processing of an image pair took about 26.5s on an AMD Athlon 64 processor (2GHz, 32-bit mode) and the mean segmentation error amounts to $\mu^{\text{class}}=0.141$. Using $d_j<0.125$ as criterion for the generation of additional training data, even 81.3% of the cells were recognised. Provided a significance level of $\alpha=0.05$ is applied, these recognition rates do not differ significantly from each other. Moreover, this modification caused a rise of the FPRs by 100%. Therefore, it was rejected.

On the whole, the recognition rates are comparable with the ones achieved regarding Sf9-cells (see Section 5.5). As the appearance of S2R+ cells imposes several additional problems, which had to be dealt with, it can be concluded that a similar approach applied to Sf9 cells would lead to an improved cell recognition. But it would require additional computation time as well.

5.7 Summary

In this chapter, an approach to the automatic recognition of unstained live Sf9 cells was introduced. As its application in conjunction with a protein localisation technique was intended, special requirements had to be fulfilled. Therefore, bright-field microscopy was employed. It enables a parallel acquisition of fluorescence micrographs without requiring a change of the principal optical path from the specimen to the camera. Furthermore, the proposed approach does not require the usage of additional dyes, which might interfere with proteins under analysis. This constitutes a major advantage in comparison to alternative techniques.

In order to recognise cells in bright-field images, information regarding their membranes is exploited. These cell membranes allow for a separation of cells from other cells and their surroundings. Therefore, cell membrane information was incorporated at all levels of the cell recognition – namely the localisation of cell candidates, their segmentation and the classification of the resulting segments (see Figure 5.2). So, the relatively low contrast of bright-field images could be compensated for.

The recognition rates given in Table 5.5 show that the developed techniques are able to recognise unstained live Sf9 cells. Here, recognition rates up to 90.1% were achieved. Furthermore, the image regions associated with each cell were determined very accurately. Hence, the proposed approach is applicable within the context of protein localisation.

However, the discussed techniques were tailored to a specific cell type – Sf9 cells. In order to evaluate whether or not they can be generalised to alternative cell types, experiments based on S2R+ cells, which exhibit different characteristics, were conducted. This entailed several adaptations. Firstly, additional DIC images had to be incorporated; like bright-field images, they allow a combination with fluorescence microscopy. Secondly, a new method to determine the relevant parameters was required. Finally, the feature set of the classification approach had to be adapted. Nonetheless, the principal techniques were transferred. Here, recognition rates up to 81.3% could be reached (see Section 5.6.3). So, it was shown that the proposed cell recognition approach can be successfully adapted to alternative cell types.

6 Protein Localisation

After having ensured that the majority of the considered cells is recognised correctly, an automatic localisation of tagged proteins became realisable. Here, only Sf9 cells were analysed. Nonetheless, in principle, an examination of S2R+ cells could be performed in a similar way.

In order to determine protein locations, an image region found by the cell recognition method is applied to a fluorescence micrograph taken simultaneously with each bright-field image. The analysis of the distribution of tagged proteins in the considered image region enables the identification of the proteins' locations. In the optimal case, each protein location causes a pattern distinct from all other locations. In practice however, the distribution patterns are not always unambiguous. Figure 6.1 depicts exemplary fluorescence micrographs illustrating the ten protein locations regarded within the scope of this thesis. They correspond to specific cell compartments or combinations thereof.

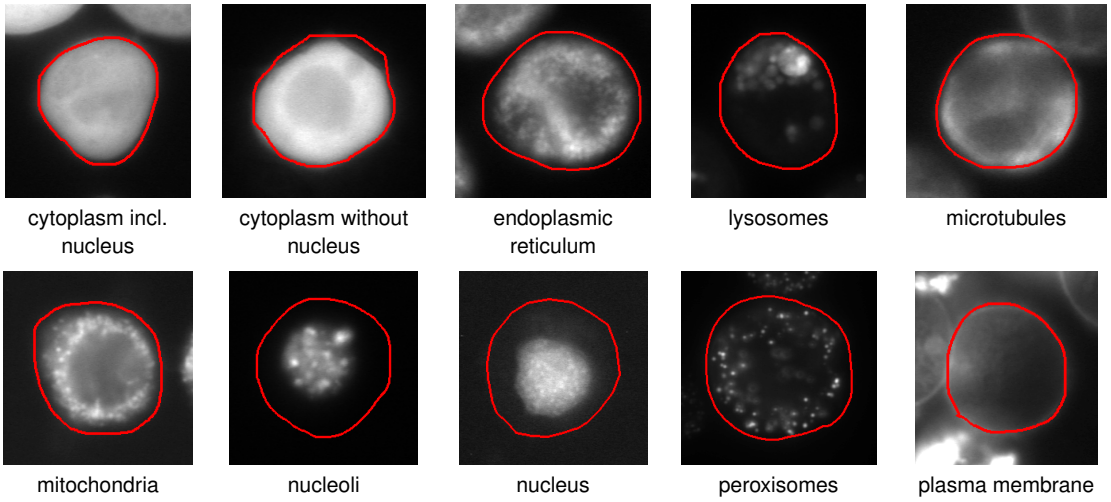


Figure 6.1: *The ten protein locations considered.* The red contours represent the surrounding cells, which were manually extracted from corresponding bright-field images by a biological expert. Some distribution patterns, e.g. for the endoplasmic reticulum and the microtubules, resemble each other very closely.

In order to solve ambiguities, alternative microscopy techniques could be applied. In particular, three-dimensional imaging techniques (see Section 4.3) would be beneficial, as they might reveal additional information. However, in order to enable a high-throughput application, time-consuming techniques, such as digital convolution and confocal laser scanning microscopy, are not very well-suited. Nevertheless, spinning disk microscopy constitutes a promising alternative for future investigations.

Since fusion proteins obtained using fluorescent proteins were only available for a subset of the considered locations, conventional dyes staining specific cell compartments had to be applied as well. Otherwise, a reasonable variety of protein distribution patterns could not have been collected. However, as the resulting fluorescence micrographs are assumed to be very similar to the corresponding protein distribution patterns, the negative influence of this procedure is limited. Table 6.1 lists the protein locations and the respective dyes.

cell compartment	dye
cytoplasm including nucleus	GFP (no fusion protein)
cytoplasm without nucleus	fusion protein (GFP)
endoplasmic reticulum	3,3'-dipentyloxacarbocyanine iodide $\left(\text{DiOC5(3)}\right)$
lysosomes	Lysosensor Green
microtubules	paclitaxel-Oregon Green
mitochondria	MitoTracker Orange
nucleoli	fusion protein (YFP)
nucleus	bisbenzimide
peroxisomes	fusion protein (GFP)
plasma membrane	FM 1-43

Table 6.1: *Applied staining methods.* The images depicting several protein localisations were obtained by constructing fusion proteins, while others were acquired using conventional dyes.

Before the proposed approach to automated protein localisation is detailed, I want to address related research results and methods known from literature (cf. Section 6.1). The actual localisation procedure is realised by classifying the observable protein distribution patterns in specific image regions, which correspond to single cells. These regions can either be obtained by means of manual extraction or determined automatically using the cell recognition approach introduced in Chapter 5.

In order to perform a classification of the regarded image regions, adequate features are computed first. They are discussed in Section 6.2. Based on these features, datasets were recorded. The process of dataset generation is described in Section 6.3. Here, an adaptation to the proposed cell recognition method was performed so that the cooperation of both approaches is facilitated, which is beneficial regarding the goal of automatic protein localisation.

As discussed in Section 5.4.3 a reduction of the feature set might be advantageous with respect to the classification task. Therefore, two applicable feature reduction methods are introduced in Section 6.4. Here, the focus lies on feature selection techniques, as they enable an easier interpretation of the resulting features in comparison to feature extraction approaches. This is intended to support biologists, who use the protein localisation procedure. With respect to cell recognition, such an interpretation has not been required.

The analysis of the classification occurred in two steps: Firstly, off-line learning was performed; i.e., a classifier was trained to distinguish between the ten considered compartments (see Section 6.5). Afterwards, an evaluation of the on-line learning capabilities occurred (see Section 6.6). Here, each classifier was trained using only a subset of the regarded cell compartments. The remaining data was employed for an additional training as it might be desired during the application of a protein localisation mechanism; in particular, if previously unknown protein distribution patterns show up.

Finally, Section 6.7 summarises the present chapter. Here, the results with respect to the approaches using a fixed and a trainable set of classes are discussed.

The methods discussed in the present chapter were published in [229]. But the investigations introduced therein are based on a subset encompassing six different protein locations to be sep-

arated. Therefore, a reduced basic[1] feature set sufficed. In order to allow for a recognition of ten locations, it had to be extended.

6.1 Related Work

As discussed in Section 3.5.3, the analysis of fluorescence micrographs is a common means for determining the subcellular location of proteins. Frequently, this is performed by the experimenters themselves, who incorporate some prior knowledge [89][96]. Nevertheless, additional tools, for instance, image editing software [203] or auxiliary images [121] are required. Unfortunately, such manual analyses are subject to the experimenters' training and experience. Furthermore, these investigations are very time consuming and barely allow for the examination of large proteomes. Therefore, there is a need for automated methods enabling high-throughput analyses.

The research on how to automatically recognise protein distribution patterns obtained by means of fluorescence microscopy has been dominated by the American biochemist Robert F. Murphy and his group. They have experimented with a multitude of different methods, in particular, microscopy techniques [94], numerical features [36][89][147], feature reduction methods [95], classifiers [17][94] and clusterers [70]. They usually consider ten different cell compartments [36][89][94][95][147] if two-dimensional images are applied. Using three-dimensional images, up to eleven different protein locations [35][94] are examined. Their early approaches regarded significantly less protein locations; for example, the technique suggested in [16] distinguishes only between four localisations and DNA. However, an additional class reflecting the amount of unknown patterns was introduced. But is has been dropped in more current research.

In recent years, the interest in developing automated methods for the subcellular localisation of proteins has increased. Christian Conrad and his colleagues proposed a feature-based machine learning approach to the high-throughput analysis of twelve protein locations in live human cells [44]. Here, the evaluation is based on two-dimensional images of single cells, which were counterstained with an additional dye. As an alternative to using numerical features, Anne Danckaert and her colleagues suggested the usage of a neural network structure, which enables the recognition of six protein locations using down-scaled images of single cells [46]. In order to achieve this goal, a stack of confocal images is searched to determine the image slice with the highest information content. This slice is passed to the neural network. Besides the cell compartments to be classified, this technique allows for the recognition of ambiguous patterns.

Other researchers have employed techniques tailored to specific locations: Peter M. Kasson and his co-researchers have proposed a technique for the classification of proteins localised at the plasma membrane [112]. David A. Schiffmann and his colleagues used counterstaining in order to measure the protein concentration in the kinetochores[2] [188]. A similar method has been recently applied by Sreevatsan Raman and his co-researchers [173]. They have analysed abnormalities of nuclear compartments called centrosomes by means of counterstaining the nucleus and exploiting radial symmetries. Furthermore, Urban Liebel suggested a simple image processing-based technique enabling large-scale screens of proteins localised in the Golgi apparatus [127].

From a biological point of view, systems allowing for the differentiation between more then 20 different protein locations would be beneficial. The team of Won-Ki Huh at the University

[1] feature set before performing any kind of feature reduction
[2] cell compartments involved into mitosis

of California, for example, assigned 75% of the yeast proteome to one of 22 distinct locations [96]. They employed a set of twelve basic location patterns. Fluorescence images of each protein were manually assigned to one of these locations. Afterwards, the categories were refined by means of co-localisation experiments. This approach indicates an upper limit for candidate protein localisation techniques; if a human expert needs additional knowledge in order to assign an image to more then twelve locations, an automatic method probably needs this information as well.

However, Shann-Ching Chen and his co-researchers investigated the application of a protein localisation approach to the data of Huh's team. This approach did not exploit information from co-localisation experiments [33]. At first glance, their results seem promising, since total accuracies of about 81% were reached. But here, it has to be taken into account that the number of available images strongly varied depending on the considered cell compartment; for example, they used 819 images of proteins located in the cytoplasm, while only ten images showing proteins in the microtubules were available. So, a mean accuracy over all cell compartments rather than the total accuracy, which measures the total amount of correctly classified images, should be considered. The average of the diagonal elements of the confusion matrix [33, Table 3] constitutes such a measure. It amounts to 43.1%, which clearly shows an extremely low level of classification accuracy regarding individual cell compartments.[3] So, the need for additional co-localisation experiments has been confirmed.

In principle, the approaches to protein localisation can be divided into groups according to several important properties: Firstly, the employed microscopy technique is critical with respect to the images to be analysed. Here, at least one of four techniques is employed. These techniques comprise basic fluorescence microscopy (see Section 4.2.6) [29][44], fluorescence microscopy in conjunction with digital deconvolution (see Section 4.3.1) [32][89][94][112], confocal laser scanning microscopy (see Section 4.3.2) [29][46][121] and spinning disk microscopy (see Section 4.3.3) [70]. The latter three yield three-dimensional stacks of images, while the basic fluorescence microscopy is based on two-dimensional micrographs. In principle, basic fluorescence microscopy enables the fastest acquisition of images. On the other hand, the three-dimensional techniques provide more detailed information, which might alleviate the protein localisation. Here, an application of the digital deconvolution has the additional drawback that the provided image details do not necessarily reflect the real appearance of biological specimens. Due to its high speed and its relatively small amount of data generated, basic fluorescence microscopy without deconvolution is particularly well-suited to high-throughput approaches.

Micrographs showing protein distribution patterns usually do not provide sufficient information regarding the surrounding cells (see Chapter 1). Therefore, additional knowledge is necessitated. Several approaches to protein localisation utilise images of single cells. These images might have been cropped manually [16][35][46][89] or computed using counterstaining [35][44][89]. In some approaches, information resulting from counterstaining is employed so as to segment the cells, e.g. by means of the watershed transform [35]. Depending on the type of protein localisation pattern, it might be applicable to cell segmentation as well; for example, if the proteins under consideration are located close to the plasma membrane, they can be used for cell segmentation [112].

Alternatively to single-cell images, recently some approaches analysing images that depict multiple cells have been proposed [93]. Their advantage consists in the fact that they do not need a

[3] As five of 20 protein location patterns were not recognised at all, the mean accuracy $\overline{\mathrm{ACC}}$ (see Section 6.5), which is used to assess the results given in my thesis, is not computable (division by zero).

cell recognition step. But they require homogeneous location patterns for all cells, which cannot be guaranteed for numerous biological experiments. Furthermore, image properties different from protein locations, e.g. the cell distribution, might influence the outcome of the analyses. Shann-Ching Chen and his co-researchers performed a comparative study, which consequently showed the superiority of the cell-based technique [33]. They pointed out the possibility of combining the result of several single-cell classifiers. In the case that homogeneous localisation patterns occur or rather the distribution of protein locations is known, the distribution of the single-cell classification results can be exploited to improve the recognition result; for example, the majority vote of all classifiers could be assigned to the complete image.

Assuming the classification is based on single-cell images, one important question has to be answered: How does the cell recognition approach affect the outcome of the protein localisation? Even if the cell segmentation is performed manually, there might be differences between several human experts, let alone automatic segmentation methods. These differences could affect the protein localisation, for instance, if numerical features are applied, which reflect the intracellular position of the tagged proteins. Unfortunately, this question has usually been left open.

In comparison to the direct application of pixel intensities, the usage of numerical features has proven advantageous for the classification of fluorescence images showing tagged proteins [36][147]. In the literature, a multitude of different feature sets has been proposed [44][89][147]: morphological features, histogram-based features, edge-related features, convex hull features, moment-based features, features based on co-occurrence matrices and wavelet features. Here, usually several types of features are required to allow for a correct classification of protein location patterns. So, if a new cell line is to be analysed, the problem consists in defining a set of feature types which reflects the location pattern adequately. Selecting only an individual feature type most probably does not suffice.

After choosing appropriate feature types, feature reduction is a beneficial preprocessing step (cf. Section 5.4.3), which is performed by numerous approaches to protein localisation [33][35][44][95][70]. In particular, the decreased dimensionality of the data may result in a higher performance of the applied classifier. Here, especially a feature selection method called *stepwise discriminant analysis* [44][95] and *genetic algorithms* [95] have proven beneficial. Therefore, I decided to use these methods as well. However, the applied genetic algorithm (cf. Section 6.4.2) differs significantly from the one employed in [95].

For the final classification, numerous classifiers have been utilised, in particular, k-nearest neighbour classifiers [91], neural networks [17][44][91][94], support vector classifiers using several kernels [33][44][91][94] as well as ensemble methods such as bagging [94] and boosting [91][94]. Here, especially the neural networks and the support vector classifiers performed very well on a great variety of datasets. In this context, the term ‘neural network’ usually refers to multilayer perceptrons [202][275, Chapter 7]; alternative neural approaches, such as radial basis function networks [202][275, Chapter 20] or ARTMAP networks (cf. Section 5.4.4) [22], are barely taken into account.

In recent years, the interest in automatically identifying distinct location patterns has risen [34][36][70][145][146]. Here, the applied techniques differ considerably from the ones used for protein localisation so far: Supervised learning mechanisms have been replaced by unsupervised ones, which do not incorporate prior knowledge on the data processed. These unsupervised learning methods summarize similar location patterns to so-called clusters. In comparison to classifiers,

the number of clusters typically exceeds the number of classes or rather the a priori considered protein locations. However, in principle, the whole process is controlled by a similarity criterion, which reflects the similarity of the regarded images in the feature space and determines the outcome. So the clustering process depends on the chosen features and the utilised similarity criterion.

In order to find relevant location patterns, Xiang Chen and his colleagues proposed a method, which distinguishes between all proteins under analysis, even if they share a common location [34]. It is applicable to single-cell images [34][36] as well as multi-cell images [70]. They use the k-means algorithm [218] to cluster the images independently of the proteins shown. Then all images showing a specific protein are analysed. A protein is associated with the cluster containing the majority (at least 33.3%) of the corresponding images. All other images (up to 66.6%) are dropped. If no cluster comprises more than 33.3% of a protein's images, the respective location pattern is discarded completely. So, stable connections between created clusters and proteins are established. In addition to the k-means algorithm, they apply hierarchical clustering [218] to the location patterns, which have not been dropped before. As the similarity criterion is crucial regarding the clusters formed, they evaluated two criteria with respect to their data: the Euclidean distance and the Mahalanobis distance. The quality of the similarity criteria is measured by the agreement of the sets of clusters yielded by both considered clusterers. The best agreement was achieved by the Euclidean distance. From my point of view, this result is not surprising, as the Mahalanobis distance is subject to a significantly higher number of parameters, which necessarily increases the variation of the clusters. However, it might better fit the underlying biological data. Unfortunately, this relation has not been sufficiently analysed. It is rather assumed that the clusters yielded by the systems with the highest agreement reflect the real structure of the data.

In contrast to the methods discussed above, the protein localisation technique introduced in this thesis aims at integrating all components required to localise proteins automatically. This encompasses a cell recognition method, a classifier realising the differentiation between a priori known location patterns and mechanisms, which allow for the recognition and incorporation of new location patterns during the application of the suggested protein localisation system.

6.2 Features Reflecting Protein Location Patterns in Sf9 Cells

The proposed approach to the evaluation of protein distribution patterns is based on three principal types of features: Firstly, features enabling a consideration of the positions of tagged proteins relative to the surrounding cells are employed. They comprise Zernike moments (cf. Section 6.2.1) and region-dependent texture features (see Section 6.2.2). The latter are related to the features utilised for the cell recognition (see Section 5.4.1). Both are invariant regarding rotations, translations and scale changes. Secondly, I decided to apply morphological features, in particular pattern spectra (cf. Section 6.2.3), which allow for an evaluation of the shape and the size of protein accumulations. Finally, general properties of the protein distributions are regarded by means of fractal features (see Section 6.2.4) and histogram-based features which are applied to the whole segment (cf. Section 6.2.5). So, characteristics less obvious than shape, size and location of protein accumulations are incorporated into the localisation procedure, for example, the heterogeneity and roughness of the image at different scales. Similar to the region-dependent texture features, the histogram-based features resemble the features employed during the cell recognition (see Section 5.4.1). The morphological, fractal and histogram-based features are invariant with respect to

changes of the orientation and the position of cells. However, the size of cell compartments is an important property. So they are not scale invariant.

All features were chosen in such a way as to yield a comprehensive view of the protein distribution patterns which might occur. Therefore, an incorporation of additional locations should be possible without any problems. Furthermore, the features are rotationally invariant and translation invariant, which constitutes crucial properties, as the orientations and image positions of the considered cells differ.

However, one important problem needs to be solved before an application of the introduced features is reasonable. Even if the fluorescence patterns of different cells resemble one another, the corresponding fluorescence magnitudes might differ significantly. Therefore, the image intensity has to be normalised with respect to each recognised cell segment. Otherwise, less intensive fluorescence patterns could not be classified correctly. Moreover, in the case that a segment encloses fluorescent parts from a neighbouring cell, the normalisation and consequently the classification might fail. Therefore, the protein localisation requires an accurate cell recognition.

6.2.1 Zernike Moments

Zernike moments are named after the Dutch physicist Frits Zernike (cf. Section 4.2.3). He employed a set of orthogonal polynomials $V_{pq}(r,\alpha)$ to investigate the characteristics of phase contrast microscopy [276]. Their general layout is shown in Equation 6.1 [117]. Due to its inventor, they are referred to as Zernike polynomials.

$$V_{pq}(r,\alpha) = R_{pq}(r)e^{iq\alpha} \qquad \text{with} \qquad i = \sqrt{-1} \tag{6.1}$$

Here, the data points are represented using polar coordinates: r denotes the radius and α the angle. The parameters p and q determine the degree of the polynomial. While p can only be set to integer values greater than or equal zero, for q negative integer values are allowed as well. But the following conditions must be fulfilled: $p-|q|$ is even and $|q| \leq p$. $R_{pq}(r)$ constitutes a radial polynomial which is defined according to Equation 6.2.

$$R_{pq}(r) = \sum_{s=0}^{\frac{p-|q|}{2}} (-1)^s \frac{(p-s)!}{s!\left(\frac{p+|q|}{2}-s\right)!\left(\frac{p-|q|}{2}-s\right)!} r^{p-2s} \tag{6.2}$$

As Zernike polynomials are complex functions, they comprise a real as well as an imaginary part. This is illustrated in Figure 6.2. Due to their dependency on the angle α, both parts are not rotationally invariant.

Based on the complex polynomials $V_{pq}(r,\alpha)$, the *Zernike moments* A_{pq} are computable. In principle, they constitute projections of an image on the Zernike polynomials. These projections can be determined according to Equation 6.3. Here $V^*_{pq}(r,\alpha)$ symbolises the complex conjugate of $V_{pq}(r,\alpha)$.

$$A_{pq} = \frac{p+1}{\pi} \sum_{0 \leq r \leq 1} \sum_{0 \leq \alpha < 2\pi} I\big(x(r,\alpha), y(r,\alpha)\big) V^*_{pq}(r,\alpha) \tag{6.3}$$

As $V_{pq}(r,\alpha)$ is a complex quantity, the moments A_{pq} are complex as well. Furthermore, they inherit the dependency on rotations. This is inappropriate regarding the subcellular localisation of

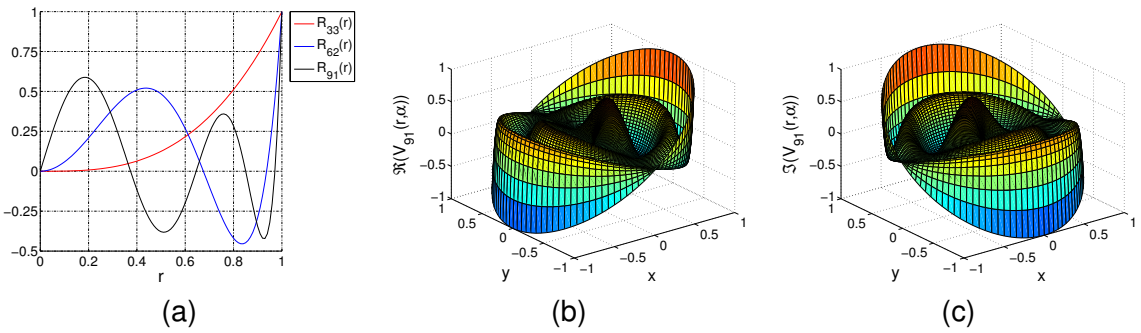


Figure 6.2: *Principle of Zernike polynomials.* The complexity of the radial polynomials $R_{pq}(r)$ varies depending on the parameters p and q (a). Their basic shape is transferred to the Zernike polynomials $V_{pq}(r, \alpha)$, where it is modified with respect to the angle α. In order to alleviate the understanding, the real part $\Re\big(V_{91}(r, \alpha)\big)$ (b) as well as the imaginary part $\Im\big(V_{91}(r, \alpha)\big)$ (c) of the exemplary polynomial $V_{91}(r, \alpha)$ have been drawn using Cartesian coordinates.

proteins, since the orientation of cells and the position of tagged proteins inside these cells are not related. But rotationally invariant features can be easily constructed from the Zernike moments, as any kind of rotation is expressible in terms of a phase shift [117]. So, the magnitude of the Zernike moments remains untouched. Thus, $|A_{pq}|$ constitutes a rotationally invariant feature, describing the position of tagged proteins within the cells in question. Here the term position refers to the relative distance from the cells' centres.

Equation 6.3 further implies that a suitable map from the Cartesian image coordinates to the polar coordinates of the unit circle is required. This map has to account for different sizes and shapes of cells. Then, the obtained features are scale invariant and translation invariant, as well. Such a map could be realised by normalising the total number of pixels that each segmented cell comprises [117]. Unfortunately, after performing this type of mapping, the maximum radius of a round cell differs from the maximal radius of a slightly ellipsoid one. As a result, either parts of ellipsoid cells would be outside the unit circle or there would be a certain distance from round cells to its boundary. Both consequences are not desirable. Therefore, I approximate each cell segment by an ellipse [65]. This yields the centre of the segment and two axes: the semiminor axis and the semimajor axis. Afterwards, every cell can be mapped into the unit circle in such a way that its boundary touches the circle's boundary regardless of its shape. So a comprehensive analysis of all cells is enabled.

Within the scope of my thesis, I usually employ Zernike moments up to order twelve. So a total of 49 moments is available. These features allow for a detailed description of the positions of proteins in cells. They are listed in [117].

6.2.2 Region-Dependent Texture Features

Although Zernike moments are known to reflect the positions of tagged proteins inside cells [36], they are not necessarily the best means of description. The computation of complex functions and factorials, in particular, causes a high computational load. So region-specific features similar to the ones applied during cell recognition (cf. Section 5.4.1) were utilised as an alternative. But in order to cover all possible positions, the regions need to be more detailed, especially inside the cells. Furthermore, the number of features was chosen in such a way as to equal the number of Zernike moments used. Therefore, the seven regions depicted in Figure 6.3 were selected.

The regions shown in Figure 6.3 are computed using the same map to polar coordinates that

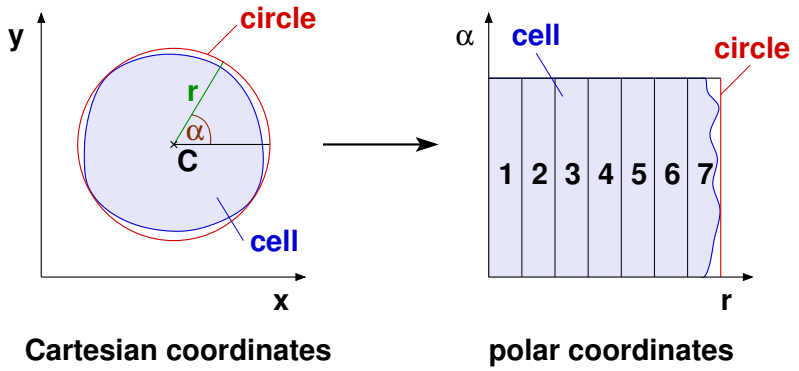


Figure 6.3: *Considered image regions.* Each cell segment is mapped to a circle with the centre C and transformed into polar coordinates (r, α). Then, seven regions of equal width are analysed. These regions correspond to disjoint tubes in the Cartesian coordinate system.

has been used for the determination of Zernike moments. So, the results are better comparable and the resulting features inherit the scale-invariancy and the translation invariancy. Moreover, the regions' shapes ensure that they are rotationally invariant as well.

Based on these regions, histogram-based features are computed. They were adopted from the cell recognition procedure introduced in Chapter 5. But in order to reach a total number of 49 features, which is intended to ensure the comparability with the Zernike moments, only seven rather than nine features are applied to each image region: the mean, the variance, the skewness, the uniformity, the smoothness, the entropy and the median (cf. Appendix B.1).

6.2.3 Granulometries and Pattern Spectra

Several cell compartments, for instance the mitochondria and the lysosomes, appear in fluorescence images as bright spots of varying size if tagged proteins are located there. Hence, a means of analysing the size and shape of these spots is required. The *granulometry*, first introduced in [139, Chapter 1], constitutes such a technique. It originates from material sciences and can be described as follows:

> *"When analysing granular materials, a granulometry is performed by sieving a sample through sieves of increasing mesh size while measuring the mass retained by each sieve."*
>
> Pierre Soille (2003) [206, page 318]

The concept of the granulometry has been transferred to digital images by applying morphological operators, in particular the morphological opening. The opening operator exhibits several properties of a sieve, as it maintains only image structures a specific structuring element fits into [206]. In order to build a granulometry, a series of openings with structuring elements of increasing size is performed. These structuring elements have to satisfy the *absorption property*; i.e., the result of two subsequent openings with different structuring elements must be equivalent to applying one opening with the larger one.

The pixels sums of the results of the openings constitute the *granulometric curve* which can be analysed by means of its discrete derivative – the *pattern spectrum* [138]. The elements of the pattern spectrum are tantamount to the fraction of image structures of specific sizes; that is, they

enable the extraction of size information without requiring a prior segmentation. Pattern spectra have been frequently applied for the texture analysis of images containing connected granular structures. Examples of its application are the analysis of images showing carotid plaque [40], coffee beans [249] and plankton [215][239]. Especially the identification of plankton, for instance diatoms[4], resembles the subcellular localisation of proteins insofar as the texture within single cells is analysed. Therefore, pattern spectra were included in the employed set of features.

Several subcellular particles and organelles, e.g. lysosomes and peroxisomes, can roughly be described as being circular. Thus I decided to employ disc-shaped structuring elements in order to enable the detection of round compartments of varying sizes. Such discs can be efficiently approximated by octagons. Octagons of arbitrary size are computable by alternating dilations with the elementary diamond[5] and the elementary square[6] [206, Chapter 11]. Since most cell organelles are rather small in comparison to the cell itself, I decided to use the first ten iterations of octagons as structuring elements. So, subcellular structures having a diameter between 3 and 21 pixels are captured.

In addition to the original concept of granulometries, shape granulometries were proposed some years ago. They enable the filtering of objects with specific shapes and are based on thinnings rather than openings [240]. As the investigated intracellular structures in the considered images differ mainly in size and not in shape, this extension has been neglected. Furthermore, there are some algorithms that increase the computational efficiency in particular for large or specially shaped structuring elements. An overview of these methods is given in [250]. Due to the comparably small size of the considered cell organelles and their mainly circular shape, these optimisations are of less importance for the protein localisation task at hand.

6.2.4 Fractal Features

Geometrical descriptions do not always suffice to describe natural structures and processes due to their inherent complexity:

> *"Why is geometry often described as 'cold' and 'dry?' One reason lies in its inability to describe the shape of a cloud, a mountain, a coastline, or a tree. Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line. "*
>
> Benoit Mandelbrot (1983) [136, page 1]

Therefore, Benoit Mandelbrot has developed a more general method for characterising these phenomena – *fractal geometry* [135][136]. Besides its relation to human perception [161], it is considered particularly well-suited for describing biological structures such as complete organisms, single cells and cytoplasmic membranes [258]. Consequently, the application of fractal features to the subcellular localisation of proteins has been investigated.

Mandelbrot's approach takes into account that complex natural shapes often possess an invariance under changes of magnification. This self-similarity as well as the irregularity of natural phenomena is measurable by means of the *fractal dimension* D^F, which might exceed the topological dimension D^T [106][251]. Several methods have been proposed for the computation of

[4] unicellular algae

[5] a pixel and its four vertical and horizontal neighbours

[6] a pixel and its eight vertical, horizontal and diagonal neighbours

the fractal dimension of two-dimensional textures. Examples are the blanket method [154][160], a technique based on the Fourier transform [161] and box counting [68][115][186][251].

Instead of the fractal dimension, the parameter $H \in [0, 1]$ is applied frequently. It enables the comparison of fractals with different topologic dimensions and is related to D^F according to Equation 6.4. In terms of images, a large value of H characterises a smooth texture and a small value a rough texture.

$$D^F = D^T + 1 - H \tag{6.4}$$

The parameter H can be obtained from images according to Equation 6.5 where $E(|\Delta I|)$ denotes the expectation of the intensity difference of two neighbouring pixels, d their distance and k a constant factor [265].

$$E(|\Delta I|) = k \cdot d^H \tag{6.5}$$

But a fractal cannot be described completely by its fractal dimension. Therefore, Mandelbrot introduced a further measure called *fractal lacunarity* Λ^F , which characterises a fractal's granularity and heterogeneity [136, Chapter 34]. Lacunarity analysis [165] has been successfully applied to several texture classification tasks such as the classification of magnetic resonance images of bone structures [274] as well as ultrasound images of the carotid plaque [40] and the liver [265]. Since in the fluorescence images under analysis the considered cell compartments are of varying size and structure, the characterisation of the heterogeneity of textures provided by the fractal lacunarity has been considered a useful feature for their differentiation.

Chung-Ming Wu and his colleagues [265] proposed a method for the computation of fractal image features. Besides texture roughness, this technique captures information about the lacunarity of an image. Because of the application of a pyramidal approach, it is especially computationally efficient. Moreover, such an analysis of different resolution levels has been proven beneficial regarding the task of protein localisation [32].

In Chung-Ming Wu's technique, the image is observed at n different resolution levels i, each of which decreases the size of the previous image by 50% in every dimension. The original image receives the index $i{=}0$. For these resolution levels, the parameters H_i are estimated, which leads to the *multiresolution fractal feature vector* (MFFV) shown in Equation 6.6.

$$\text{MFFV} = \big(H_0, H_1, H_2, \ldots, H_{n-1}\big) \tag{6.6}$$

In order to determine H_i for a specific resolution level i, the intensity difference vector IDV_i has to be computed first (see Equation 6.7). It contains the mean intensity differences for pixels up to a maximal distance $d^{\max}$.

$$\text{IDV}_i = \big(\text{id}_i(1), \text{id}_i(2), \text{id}_i(3), \ldots, \text{id}_i(d^{\max})\big) \tag{6.7}$$

The mean intensity difference $\text{id}_i(j)$ for pixels of distance j can be determined according to Equation 6.8.

$$\text{id}_i(j) = \frac{\sum\limits_{\forall (x,y) \in \mathcal{X}} \big|I(x,y) - I(x+j,y)\big| + \sum\limits_{\forall (x,y) \in \mathcal{Y}} \big|I(x,y) - I(x,y+j)\big|}{|\mathcal{X}| + |\mathcal{Y}|} \tag{6.8}$$

In contrast to the approach of Chung-Ming Wu and his co-researchers, the task at hand requires that only pixels of a specific segment are considered instead of the whole image. Therefore, the sets $\mathcal{X}$ and $\mathcal{Y}$ were introduced, in order to handle arbitrarily shaped image regions. They encompass all points which possess a neighbour that lies in the current segment and has a distance of j pixels in the direction of increasing x and y, respectively. In the case of processing a complete image, Equation 6.8 is equivalent to the original definition. Due to Equation 6.5, the value of each H_i corresponds to the slope of the regression line of $\log\left(\mathrm{id}_i(j)\right)$ on $\log(j)$. It can be computed using well-established methods [57, Chapter 3].

The maximal resolution level n as well as the maximal distance $d^{\max}$ were chosen in such a way as to allow for a comprehensive analysis of the considered cells: n was set to 5 and $d^{\max}$ to 10. As the maximum cell diameter of the considered Sf9 cells amounts to 322 pixels, higher resolution levels were not reasonable ($322/2^5 \approx 10$).

6.2.5 Histogram-Based Features

In order to extend the description of textural properties, nine additional histogram-based features were utilised. In contrast to the features discussed in Section 6.2.2, the complete cell is analysed without performing any coordinate transformation. The regarded features are the mean, the variance, the skewness, the uniformity, the smoothness, the entropy, the 5th percentile, the 50th percentile (median) and the 95th percentile (cf. Appendix B.1).

6.3 Generation of Datasets

The required training and validation datasets were acquired in a similar way like the ones employed during the cell recognition (cf. Section 5.4.2): Firstly, cells masks, which had been manually extracted from bright-field images by a biological expert, were utilised. But in contrast to the cell masks applied to cell recognition, each cell mask was associated with a fluorescence micrograph showing a specific protein location pattern (cf. Figure 6.1). Secondly, an automatic generation of additional training data occurred using the localisation and segmentation procedures described in Section 5.3. The application of various values for the relevant parameters, in particular, the energy weights of the greedy snakes, enabled the computation of segments which resemble the manually determined cells. Therefore, the corresponding regions of the respective fluorescence micrographs were evaluated to localise tagged proteins as well. These segments are more likely to occur than the cell masks if the proposed protein localisation technique is applied in conjunction with the cell recognition method introduced in Chapter 5. So, the protein localisation technique was adapted to the utilised cell recognition approach. In addition, the number of training samples was increased, which alleviates the classification task. Otherwise, the number of training samples might not have been sufficient.

The choice of the similarity criteria for accepting generated segments as cells occurred in such a way as to be compatible with the cell recognition approach; i.e., $E_j^{\text{seg}} = d_j$ and $E_j^{\text{seg}} = \max\left(A_j, d_j\right)$ were employed. Segments leading to $E_j^{\text{seg}} < 0.1$ were considered as cells. But non-cells were not recorded. These segments, which are required to recognise cells correctly, do not have any meaning with respect to protein localisation. Therefore, neither automatically nor manually determined segments showing objects other than cells were employed here.

6.4 Feature Reduction

In contrast to the proposed cell recognition approach, which employed mainly feature extraction techniques (cf. Section 5.4.3), here, the application of feature selection methods is favoured; they enable an easier interpretation of the resulting features, which might facilitate the investigation of biological questions. Such a kind of interpretation was not necessary regarding the classification of obtained segments during the process of cell recognition, as these segments are the result of the automatic localisation and segmentation techniques rather than originating from biologically relevant structures.

The first feature selection method applied consists in stepwise discriminant analysis (SDA) (see Section 6.4.1). It chooses a set of features depending on statistical properties of the data. The used classifier is not taken into account; therefore, it constitutes a filter approach. From the literature, it is known that the stepwise discriminant analysis is very well-suited for selecting features in the context of protein localisation (cf. Section 6.1) [44][95]. Although no information regarding the usage of Sf9 cells was available, I expected excellent results from an application of the SDA. Thus, I decided to utilise it as well.

In addition to the SDA, genetic algorithms have proven beneficial for choosing relevant features in protein localisation tasks [95]. They constitute wrappers, as the classification accuracy of the utilised multi-class SVM is incorporated in the fitness function. Inspired by the excellent results, I developed a similar approach tailored to the applied classifiers – the simplified ARTMAP networks. It is introduced in Section 6.4.2. However, as genetic algorithms can be applied to solve general optimisation problems, its usage for filter approaches is possible as well; for example, Alejandro Sierra and Alejandro Echeverría proposed such a technique, which they refer to as *evolutionary discriminant analysis* (EDA) [201]. But, since with the SDA a filter technique has already been considered, I wanted to focus on wrappers as an alternative method.

6.4.1 Stepwise Discriminant Analysis

The *stepwise discriminant analysis* (SDA) introduced by the American mathematician Robert I. Jennrich in 1977 [104] is a method for the selection of an optimal subset of features by optimising the separation of samples $\underline{x}$ from different classes. The applied criterion is similar to the Fisher criterion [64] usually used for the *linear discriminant analysis* (LDA) [63, Chapter 9]: This Fisher criterion, named after the English statistician and geneticist Sir Ronald Aylmer Fisher, maximises the ratio of the *between-class scatter* and the *within-class scatter* represented by the matrices $\underline{S}_B(\underline{x})$ and $\underline{S}_W(\underline{x})$, respectively (see Equation 6.9) [82, Chapter 6][63, Chapter 9].

$$J_1(\underline{x}) = \mathrm{tr}\big(\underline{S}_W(\underline{x})^{-1}\underline{S}_B(\underline{x})\big) \tag{6.9}$$

The *total scatter matrix* $\underline{S}_T(\underline{x})$ can be obtained from $\underline{S}_B(\underline{x})$ and $\underline{S}_W(\underline{x})$ according to Equation 6.10.

$$\underline{S}_T(\underline{x}) = \underline{S}_W(\underline{x}) + \underline{S}_B(\underline{x}) \tag{6.10}$$

Alternatively to $J_1(\underline{x})$, the criterion $J_2(\underline{x})$ shown by Equation 6.11 can be utilised for the LDA. Here, the within-class scatter is minimised with respect to the total scatter. Both criteria are inti-

mately related. $J_2(\underline{x})$ is also called Wilks' Λ-criterion.

$$J_2(\underline{x}) = \Lambda(\underline{x}) = \frac{|\underline{S}_W(\underline{x})|}{|\underline{S}_T(\underline{x})|} = \frac{|\underline{S}_W(\underline{x})|}{|\underline{S}_W(\underline{x}) + \underline{S}_B(\underline{x})|} = \frac{1}{|\underline{I} + \underline{S}_W(\underline{x})^{-1}\underline{S}_B(\underline{x})|} \tag{6.11}$$

$\Lambda(\underline{x})$ has values from the interval $[0, 1]$ where zero indicates a poor and one an excellent separation of the considered classes. The advantage of using $\Lambda(\underline{x})$ consists in the possibility to incorporate a multiplicative increment resulting from adding or removing a feature x^*, which is called a partial Λ-statistic. The Indian statistician Calyampudi R. Rao showed in 1965 that such a Λ-distribution can be approximated by the more common F-statistic [174, Chapter 8c]. Therefore, the SDA computes in each step the significance of the change of $\Lambda(\underline{x})$ caused by the removal of a used feature as well as the insertion of a new one. The corresponding statistics are called F-to-remove and F-to-enter. If the change is significant, i.e. if the corresponding value is larger than given thresholds F_{out} and F_{in}, respectively, the modification of the feature set is performed. In order to maximise the outcome of the operation, x^* is chosen so as to cause a maximal change. If no feature can be included or deleted, the procedure is terminated.

Thus, the SDA computes a subset of relevant features by selecting single features which best fit the already determined set (F-to-enter). As the relations of the selected features might change due to the insertion of a new one, single features can be rejected after they were selected (F-to-remove). This strategy has proven beneficial in comparison to methods which solely include or solely reject features [82, Chapter 6].

However, several drawbacks of such stepwise methods in general have been reported in the literature [222]. Firstly, the resulting feature set need not necessarily be the optimal feature set of a given size. Secondly, there is a strong influence of noise. In addition, the application of the F-test imposes a crucial restriction: the features must constitute normal random variables [19, Chapter 5]. This cannot be guaranteed for all kinds of features. So, the application of the SDA should always be questioned and reviewed, for instance by means of cross-validation [62].

6.4.2 Usage of a Genetic Algorithm

There are several known approaches to feature reduction using genetic algorithms in a wrapper framework. Here, the selected feature set is usually represented as a bit string [95][176][270]; i.e., each feature is either present or absent. Additionally, a modification, such as a linear scaling [116][176] or a rotation [116], of the feature vectors is frequently performed. So, the classification accuracy can be increased. As classifiers, for instance SVMs [95], k-nearest-neighbour classifiers [116][176] and neural networks [270] have been applied. Hence, this approach seemed to be well-suited to construct a feature selection method based on simplified ARTMAP networks.

Although frequently utilised [95][176][270], I did not want to represent sets of select features as binary strings in the genome, as this causes discontinuities in the objective function, which impede the optimisation performed by the genetic algorithm [176]. Even if the inclusion of an individual feature is determined by multiple mask bits, the majority of which is responsible for rejecting or accepting a feature, the discontinuities remain. Therefore, I focused on the usage of numerical weights w_j, which linearly scale the respective features x_j. So, they control a classifier's ability to separate classes along the corresponding axes [176]; if the weight is high, the separation is alleviated and vice versa.

The components of the weight vector $\underline{w}$ as well as the parameters ρ and τ are evolved by the genetic algorithm. Such an inclusion of classifier-specific parameters into the genome has been reported to be advantageous [116][176]. All quantities to be optimised are scaled to fit in the interval $[0, 1]$ (cf. Section 5.6.1). In order to handle slight differences in the fitness values of the population, I utilised rank-based selection computed by means of non-deterministic linear sampling [59, Chapter 8]. Furthermore, arithmetic crossover and mutation for continuous-valued genes are employed [59, Chapter 9]. According to preliminary experiments, the crossover probability was set to 0.1 and genomic changes caused by the mutation operator were sampled from the normal distribution $\mathcal{N}\left(0, 0.025^2\right)$ (cf. Section 5.6.1). In order to reliably obtain good solutions to the problem at hand, the genetic algorithm is run over 100 generations using 100 individuals each. These values had proven sufficient in preceding analyses.

The fitness $f(\underline{g}_i)$ corresponds to the cross-validation accuracy ACC of the classifier with the index i that has been diminished by a punishment for large values of τ and high weights w_j. These punishments are scaled by the constants c_w and c_τ, respectively (see Equation 6.12).

$$f(\underline{g}_i) = \mathrm{ACC} - c_w \cdot \frac{1}{p} \sum_{j=0}^{p} w_j - c_\tau \cdot \tau \tag{6.12}$$

So, only the weights of features which are important for obtaining a high total accuracy receive high values and the considered subspace is reduced. After each run of the genetic algorithm, all weight vectors are normalised in such a way that their maximum component equals one in order to enable the usage of a maximum fraction of the input space $\left(\forall j \in \{1, \cdots, p\}\colon x_j \in [0, 1]\right)$ and to avoid multiple solutions of the optimisation function resulting from scaling. By considering the weights of the final generation, conclusions about the relevance of features can be drawn, as these individuals are adapted to the task at hand.

The first results using this approach were obtained using six different protein locations and a basic feature set comprising only 64 features [229]. Based on the resulting feature sets, here, a maximum cross-validation accuracy of 92% was reached. This accuracy did not diminish even if only 19.2 features were employed on average. But two problems remained: Firstly, the computation of the genetic algorithm is rather tedious, in particular, as five-fold cross-validation is performed to facilitate an objective evaluation of the genetic algorithm. The computation of one of the five runs takes about 30 hours using two Dual Core AMD Opteron processors operating at 2GHz. So, although four individuals are processed in parallel, the computational load is rather high. Secondly, although the features are linearly scaled according to their relevance, the feature space has constantly 64 dimensions; but several classifiers, such as the applied simplified ARTMAP networks (cf. Section 5.4.4), achieve better results using feature spaces with an appropriately decreased dimensionality.

These problems can be avoided by controlling the incorporation of particular features depending on a probability related to their weights; i.e., features possessing high weights are included very often, while features with low weights are mainly neglected. The resulting decrease of the feature space's dimensionality causes a considerable acceleration of the algorithm and might result in a higher degree of accuracy. Furthermore, in order to increase the evolutionary pressure to diminish the weights, I introduced a weight modification function $f_m(w_j)$:

$$f_m(w_j) = w_j^\kappa \qquad \text{with} \qquad \kappa > 1 \tag{6.13}$$

This function is monotone increasing in the considered interval $[0, 1]$ similar to the identity function applied before. Instead of scaling the features using w_j, they are now multiplied by $f_m(w_j)$. The probability-dependent usage of features is also controlled by $f_m(w_j)$; it equals the probability for accepting a feature. Moreover, the fitness function has been adapted:

$$f(\underline{g}_i) = \mathrm{ACC} - c_w \cdot \frac{1}{p} \sum_{j=0}^{p} f_m(w_j) - c_\tau \cdot \tau \tag{6.14}$$

As the weight modification exponent κ is usually set to 5, the weights are decreased in such a way that only very high values lead to an inclusion of the respective feature, weights smaller than 0.5 are only accepted with a probability less than 3.2%. Additionally, the evolutionary mechanisms are supported in diminishing unnecessary weights, since smaller changes are more effective if the weights are high (cf. Equation 6.14).

Besides the changes introduced above, an additional evolutionary operator is employed – *elitism* [59, Chapter 8]. It guarantees that the best individuals of each generation are passed to the next generation unchanged. So, excellent solutions to the task under consideration are maintained. However, as elitism diminishes the diversity of new populations, I decided to keep only a small fraction of the individuals of each generation (5%).

This modified genetic algorithm accelerates the feature selection: The obtained accuracy of the evolved classifiers and the number of necessary features are comparable. But the required processing time is significantly reduced. Using two Dual Core AMD Opteron processors operating at 2GHz, one trail takes about 13 hours rather than 30 hours. Therefore, only the enhanced version of the genetic algorithm is made use of. This is particularly important, since the size of the feature set has been increased in order to enable the localisation of proteins in ten instead of six cell compartments.

The *total accuracy* ACC, which reflects the amount of correctly classified patterns, has a crucial drawback regarding the evaluation of the classifiers: Insufficient results with respect to some protein locations might be balanced by others. So especially locations for which only a few training samples are available could be incorrectly classified. Therefore, I substituted the total accuracy with the *mean accuracy* $\overline{\mathrm{ACC}}$ that denotes the mean amount of correctly classified samples averaged over all regarded cell compartments. So, the fact that different amounts of training samples are available for the considered protein locations is accounted for.

However, using the arithmetic mean, the correct recognition of nine out of ten locations would still lead to a mean accuracy of at least 90%, even if 99% of the remaining protein distribution patterns were classified incorrectly. In contrast, the harmonic mean would amount to only 10%, since it punishes large differences between the values to be averaged. Therefore, as an equally correct classification of all ten compartments is intended, the harmonic mean [19, p. 241] is utilised rather than the arithmetic mean (see Equation 6.15). Here, ACC_i denotes the total accuracy of all samples of protein location i.

$$\overline{\mathrm{ACC}} = \frac{10}{\frac{1}{\mathrm{ACC}_1} + \frac{1}{\mathrm{ACC}_2} + \ldots + \frac{1}{\mathrm{ACC}_{10}}} \tag{6.15}$$

The harmonic mean is always smaller or equal to the arithmetic mean: The larger the difference between individual accuracies ACC_i, the more it decreases in comparison to the arithmetic mean. So, similar recognition results with respect to all regarded protein locations are ensured.

The final fitness function, which results from these conclusions, is given in Equation 6.16.

$$f(\underline{g}_i) = \overline{\mathrm{ACC}} - c_w \cdot \frac{1}{p} \sum_{j=0}^{p} f_m(w_j) - c_\tau \cdot \tau \tag{6.16}$$

Even if this genetic algorithm-based approach requires considerably more processing time than the SDA, its application might be beneficial. In particular, it allows for an incorporation of the employed classifier, which results in the selection of features that enable high accuracies. In contrast to the SDA, here, no assumptions regarding the features' statistical properties are made.

6.5 Protein Localisation Using a Fixed Set of Cell Compartments

The actual protein localisation is performed by classifying observed protein distribution patterns in classes corresponding to protein locations. These locations describe the cell organelles or compartments the proteins can be found in. If the relevant protein locations that might occur are known in advance, a fixed set of classes can be employed. This type of application is considered in the current section. Here, the ten cell locations shown in Figure 6.1 are utilised. The respective numbers of cells masks, which had been manually extracted from corresponding bright-field images by a biological expert, are summarised in Table 6.2.

cell compartment	cell masks
cytoplasm including nucleus	144
cytoplasm without nucleus	56
endoplasmic reticulum	142
lysosomes	222
microtubules	102
mitochondria	268
nucleoli	74
nucleus	150
peroxisomes	71
plasma membrane	97

Table 6.2: *Numbers of cell masks for the regarded protein locations.* For each cell compartment a certain number of cell masks was extracted from bright-field images. But in order to realise a protein localisation, the analysis of fluorescence micrographs must be performed. Therefore, each cell mask is associated with a corresponding region in a fluorescence micrograph that depicts a protein distribution pattern characteristic for this compartment.

These cell masks were associated with protein location patterns from corresponding fluorescence micrographs. Based on these fluorescence micrographs, datasets required for training and testing the applied classifiers could be computed. The number of samples equals the number of cell masks. So, a total of 1326 samples was available.

In order to characterise the protein distribution patterns, the features introduced in Section 6.2 were utilised. Based on them, a composition of three different feature sets occurred. They are referred to as feature set (a), (b) and (c), respectively. All of them comprise pattern spectra (see Section 6.2.3), fractal features (see Section 6.2.4) and histogram-based features (see Section 6.2.5). In addition, feature set (a) encompasses Zernike moments and feature set (b) region-dependent texture features resulting in a total of 73 basic features each. Feature set (c) comprises both Zernike moments and region-dependent texture features. As a result, it consists of 122 features.

The three feature sets were chosen in such a way as to enable a comparison of the features reflecting the position of proteins relative to the surrounding cells; in particular, Zernike moments and region-dependent texture features. As the computational load for computing Zernike moments

is relatively high, a substitution with less computationally demanding features would be beneficial. Using an AMD Athlon 64 processor (2GHz, 32-bit mode), the mean time for computing feature set (a), (b) and (c) for one of the 1326 cell masks amounts to 4.39s, 2.14s and 4.43s respectively. So, the suggested region-dependent texture features are a promising alternative to Zernike moments.

However, 1326 samples are probably too small a set to train a classifier to distinguish between ten classes using 73 or even 122 features. Therefore, additional samples were generated using the cell localisation and segmentation procedures introduced in Chapter 5. Here, automatically segmented image regions are associated and compared with the cell masks using the similarity criteria $E_j^{\text{seg}}=\max\left(A_j, d_j\right)$ and $E_j^{\text{seg}}=d_j$ (see Section 6.3). If the image regions sufficiently resemble the cell masks $\left(E_j^{\text{seg}}<0.1\right)$, they are utilised as training samples as well. So additional 4213 and 12015 samples could be generated using $E_j^{\text{seg}}=\max\left(A_j, d_j\right)$ and $E_j^{\text{seg}}=d_j$, respectively. Besides increasing the amount of training samples, this method causes an adaptation of the protein localisation approach to the employed cell recognition procedure, which facilitates a fully automated protein localisation. However, all evaluations concerning the quality of the protein localisation are carried out considering only the manually acquired samples, as their biological relevance is higher (cf. Section 5.4.5). Nevertheless, the results regarding the automatically generated patterns are usually slightly better, which might by a result of their noticeably higher number.

Based on the available training samples, basic datasets were computed. They encompass the complete feature vectors of the considered image regions. In order to enable an appropriate validation and testing of the classification results, the basic datasets were partitioned in ten disjoint subsets. The splitting was performed in such a way that no dataset comprises samples from images of cells used for another subset. Then the datasets were arranged in five groups consisting of eight datasets for training and two datasets for testing each (see Figure 6.4). The test sets of all groups are disjoint, enabling five-fold cross-validation. In comparison to employing a fixed test dataset, this procedure requires less samples. Therefore, it was favoured.

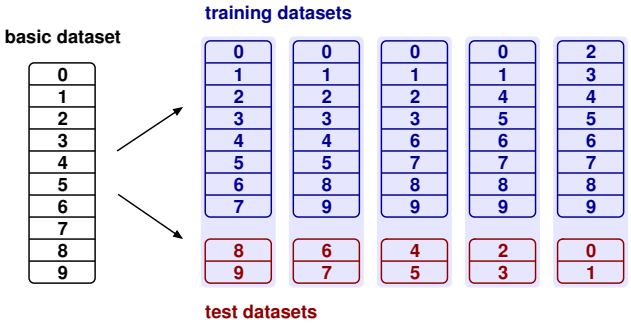


Figure 6.4: *Generation of datasets required for cross-validation.* The basic dataset is split in ten subsets. These subsets are arranged in five groups. Each group comprises eight training datasets as well as two test datasets. The test datasets of the five groups are disjoint.

The classification is performed by the simplified ARTMAP neural networks – the SFAM and the SHAM – introduced in Section 5.4.4. The support vector classifiers that were also analysed concerning the rejection of non-cell image regions are not applied here. Since at least the SFAM achieves comparable results (cf. Section 5.5), their application is not required. Furthermore, the simplified ARTMAP neural networks can be trained on-line and enable the detection of unknown[7] samples, which is crucial with respect to the desired ability of incorporating new protein location patterns into the trained system (see Section 6.6). Moreover, it allows for a direct application to multi-class classification problems. But in principle, extensions of the basic SVCs introduced in

[7] controlled by the threshold τ

Section 5.4.4 would enable their application to the task at hand if necessary [27][53][194, Chapter 7][213].

Based on the groups of datasets shown in Figure 6.4, the classifiers and feature sets were contrasted. Furthermore, similar datasets were created using feature sets reduced by the methods discussed in Section 6.4. The respective results are detailed in sections 6.5.1 to 6.5.3. Finally, Section 6.5.4 compares several important properties of the feature reduction methods used.

6.5.1 Classifying Protein Locations Based on Unreduced Feature Sets

In order to contrast the classification results, the total accuracy ACC, which reflects the amount of correctly classified patterns, was utilised. Additionally, the mean accuracy $\overline{\mathrm{ACC}}$ was employed. It denotes the mean number of correctly classified samples averaged over all regarded cell compartments (cf. Section 6.4.2). So, the fact that different amounts of training samples were available for the considered protein locations was accounted for.

The mean accuracy was further applied so as to optimise the networks' parameters ρ and τ during the training step by means of eight-fold cross-validation. Here, the networks with the highest mean accuracies with respect to the considered eight training sets were chosen and applied to the corresponding two test sets. The final results were averaged over the five groups of two test sets.

Table 6.3 summarises the classification results obtained using the unreduced feature sets. It compares both accuracy values – ACC and $\overline{\mathrm{ACC}}$ – of the networks. Furthermore, the confidence intervals of the total accuracies are given. They were computed using a confidence level of 0.95 (cf. Appendix C.3).

	feature	$E_j^{\mathrm{seg}} = \max\left(A_j, d_j\right)$			$E_j^{\mathrm{seg}} = d_j$		
classifier	set	$\overline{\mathrm{ACC}}$	ACC	conf. interval	$\overline{\mathrm{ACC}}$	ACC	conf. interval
SFAM	(a)	0.755	0.797	[0.775,0.819]	0.777	0.798	[0.776,0.820]
	(b)	0.807	0.826	[0.806,0.846]	0.818	0.833	[0.813,0.853]
	(c)	0.747	0.790	[0.768,0.812]	0.790	0.817	[0.796,0.838]
SHAM	(a)	0.707	0.750	[0.726,0.773]	0.715	0.744	[0.721,0.768]
	(b)	0.794	0.814	[0.793,0.835]	0.799	0.816	[0.795,0.837]
	(c)	0.727	0.768	[0.745,0.791]	0.735	0.769	[0.746,0.792]

Table 6.3: *Classification accuracies without performing feature reduction.* Both neural networks reached comparably high mean accuracies ($\overline{\mathrm{ACC}}$) and total accuracies (ACC) using both similarity criteria E_j^{seg} for the automatic generation of training samples. Moreover, both accuracies are affected by the basic training set used. The best results were obtained using set (b). In the case of the SHAM, these changes are even statistically significant (α=0.05).

Table 6.3 shows that the proposed methods are able to distinguish between the ten considered protein locations with a total accuracy up to 83.3%. Here, both classifiers attained similar results. The mean accuracy differs only slightly from the total accuracy. This proves that all compartments are classified sufficiently correct. The confusion matrix of the best network $\left(\text{SFAM, feature set (b), } E_j^{\mathrm{seg}}{=}d_j\right)$ is given by Table C.5 in Appendix C.4.

The influence of the similarity criterion applied to the generation of additional training samples is rather limited. Nevertheless, it indicates that the proposed protein localisation approach can be employed in conjunction with the cell recognition technique introduced in Chapter 5, since the results of the manually and automatically obtained samples are not contradictory.

The utilised feature sets appear to affect the classification, particularly, if the SHAM is applied. The best results were achieved using feature set (b), which encompasses the region-dependent texture features. So, besides decreasing the computational load, these features alleviate the classification in comparison to feature set (a). Although feature set (c) contains all features from feature set (b), the corresponding classification results are worse. This is likely to be caused by the higher dimensionality of the input space.[8] Therefore, I assumed that a feature reduction step could increase the accuracy. This assumption is further supported by the classification results obtained regarding the cell recognition (see Section 5.4.5). Consequently, I investigated the effects of reduced feature sets on the classification.

6.5.2 Classification Using Feature Sets Reduced by Means of the SDA

The stepwise discriminant analysis is a feature reduction method, which successively selects and rejects features based on their statistical properties. It has been proven to yield excellent results within the context of protein localisation [36][95]. Therefore, I decided to employ it for feature reduction. In order to achieve comparable results, I applied the procedure `STEPDISC` of the software package SAS/STAT [187, Chapter 67].

The maximum number of processing steps was limited so as to enable an investigation of the benefit of the select features using increasing feature set sizes. Since features which have been included in the reduced set are rarely rejected, the number of processing steps roughly corresponds to the size of the final feature set. The remaining parameters were set to their default values: $F_{\text{out}}=0.15$ and $F_{\text{in}}=0.15$. Such moderate values have been shown to perform very well for most applications [187, Chapter 67] including protein localisation [34].

The analysis was performed in a similar way to the analysis concerning the unreduced feature sets: The systems with the highest mean accuracies (eight-fold cross-validation) regarding the respective eight training sets were chosen and applied to the corresponding two test sets. Then, the results were averaged over the five test set groups.

Similar to the evaluation using the complete feature sets, the classification results are contrasted by means of the total accuracies and the mean accuracies. But here, the best values over all feature set sizes were chosen. In particular, the system yielding the highest mean accuracy was selected. In addition, to this mean accuracy, the corresponding total accuracy and the smallest feature set size p^*, which does not lead to a statistically significant[9] decrease of the total accuracy, is given (see Table 6.4).

Table 6.4 clearly shows that the feature reduction led to a significant improvement in comparison to the unreduced feature sets (cf. Table 6.3). This is reflected by the confusion matrices as well (see Table C.6).

Feature set (c) comprises Zernike moments and region-dependent texture features. During the process of feature selection, features of both types were chosen. Here, in particular the combined usage of the Zernike moment A_{40} and the uniformity of region 6 appear beneficial. But there is not a significant difference between the results using the feature sets (a), (b) and (c), although set (a), which does not include region-dependent texture features, seems to be slightly less beneficial. Therefore, it can be concluded that each feature type suffices to enable a correct protein localisation.

[8] The basic feature sets (a) and (b) comprise 73 features, whereas feature set (c) encompasses 122 features.

[9] significance level: $\alpha=0.05$

classifier	feature set	$E_j^{\text{seg}} = \max\left(A_j, d_j\right)$ max. $\overline{\text{ACC}}$	ACC	p^*	$E_j^{\text{seg}} = d_j$ max. $\overline{\text{ACC}}$	ACC	p^*
SFAM	(a)	0.860	0.865	6	0.860	0.873	7
	(b)	0.868	0.878	9	0.878	0.891	9
	(c)	0.866	0.881	8	0.877	0.894	7
SHAM	(a)	0.853	0.869	7	0.853	0.860	7
	(b)	0.869	0.873	9	0.879	0.890	9
	(c)	0.867	0.879	8	0.871	0.888	7

Table 6.4: *Classification accuracy using the SDA*. The feature reduction has considerably improved the accuracies of all systems in comparison to the unreduced feature sets (cf. Table 6.3). Furthermore, the size of the feature sets can be decreased dramatically until a statistically significant reduction of the accuracies occurs. The corresponding feature numbers are denoted by p^*.

Moreover, there is no considerable difference between the results of the SFAM and the SHAM as well. Thus, in terms of the accuracy, all systems are equally convenient. As a result, the classifier can be selected, which is less computationally demanding. Although, the SHAM utilises smaller weight vectors, it needs an amount of memory comparable to the SFAM, since it requires more neurons in order to obtain a high level of classification accuracy. For example, using $E_j^{\text{seg}}{=}d_j$ and feature set (b), the SFAM applies 1,131.9 nodes, whereas the SHAM employs more than 2,023.5 neurons, on average. So, both classifiers are equally well-suited regarding the task at hand.

As with the unreduced feature sets, the similarity criterion does not noticeably affect the quality of the protein localisation. Hence, $E_j^{\text{seg}}{=}d_j$, which has been proven advantageous for the cell recognition (cf. Section 5.5), is applicable to generating additional protein distribution patterns. So the protein localisation can be adapted to the cell recognition approach suggested in Chapter 5.

Since the number of features is very small, it might be of interest to have a closer look on them. The SFAM using feature set (b) and $E_j^{\text{seg}}{=}d_j$ requires only nine of 73 features in order to achieve an accuracy that does not differ significantly from the maximum accuracy. But as all results are averages over the five groups of datasets, the individual features selected might differ with respect to each group. Interestingly, six of nine features remain constant regarding the chosen system. They are listed in the following:

(i) the entropy of region 1

(ii) the uniformity of region 6

(iii) the variance of region 7

(iv) the third value of the pattern spectrum

(v) the entropy of the complete cell

(vi) the median of the complete cell

Three more features were selected using four of the five groups of datasets:

(vii) the fractal feature H_3

(viii) the sixth value of the pattern spectrum

(ix) the first value of the pattern spectrum

These nine features originate from all of the different feature types available in feature set (b): region-dependent texture features, pattern spectra, fractal features and histogram-based features. This indicates that the chosen feature types reflect different properties of the considered protein distribution patterns, which was intended during the design of the basic feature sets (a), (b) and (c). So, a comprehensive description and an accurate discrimination of different distribution patterns become possible.

6.5.3 Classification Using Feature Sets Reduced Using the Genetic Algorithm

Besides the SDA, I employed the genetic algorithm (cf. Section 6.4.2) so as to reduce the size of the employed feature set. In [95], Kai Huang and his colleagues have shown that genetic algorithms are well-suited for feature reduction in the context of protein localisation. Therefore, I developed a genetic algorithm based on simplified ARTMAP networks, which are utilised for classifying protein location patterns within the scope of my thesis. Here, the following fitness function is applied:

$$f(\underline{g}_i) = \overline{\mathrm{ACC}} - c_w \cdot \frac{1}{p} \sum_{j=0}^{p} f_m(w_j) - c_\tau \cdot \tau \tag{6.17}$$

The required parameters c_w and c_τ were determined in preceding experiments. Here, the following values have proven advantageous: c_w=0.1 and c_τ=0.02. Both constants were chosen to be rather low, which results in a dominance of the mean accuracy in the fitness function. So, networks that classify correctly are preferred. However, the lower the weights are, the higher is the fitness. So a feature reduction becomes possible. The last term of the fitness function, which is scaled by c_τ, has almost no influence. But it ensures that in the case of two individuals or rather classifiers that achieve a similar accuracy with a comparable set of features, the one using the lower value for τ is favoured. Through this, the rejection of unknown samples is facilitated.

The actual features to be utilised were derived from the individual of the final generation that had the highest fitness: The features were ordered according to this individual's weights. Then a threshold τ_w was applied to reject all features j for which $f_m(w_j)$ is smaller than τ_w. By varying τ_w, the size of the feature set can be controlled. If τ_w equals one, all features are employed; if it equals zero, the feature set solely comprises one feature.

Similar to the SDA, the evaluation of each system was performed using five-fold cross-validation; i.e., the genetic algorithm was applied to each of the five groups of training datasets (see Figure 6.4). So, five sets of weights could be determined. Depending on τ_w, which was iterated in the interval $[0, 1]$, then, five classifiers were trained with the respective eight training datasets. The test occurred using the corresponding two test datasets. Here, the accuracies were averaged. Therefore, the given results given are mean values over the five runs of the genetic algorithm using a fixed value for τ_w. The accuracies required to determine the fitness of the individuals and to choose appropriate parameter settings for the classifiers were computed by means of eight-fold cross-validation using the respective training datasets.

In Table 6.5, the classification results are contrasted by means of the total accuracies and the mean accuracies. In particular, the results of the classifier achieving the highest mean accuracy are given. Furthermore, the minimum number of features p^*, which led to a total accuracy that

did not differ significantly[10] from the maximum value is shown. In contrast to the results above, here only the similarity criterion $E_j^{\text{seg}}{=}d_j$ was employed for the automatic generation of training patterns, since the computation of the genetic algorithm is very time-consuming. But as the choice of the similarity criterion does not seem to have a considerable impact on the classification (see Tables 6.3 and 6.4), the missing results would most probably not lead to any additional insights.

	SFAM			SHAM		
feature set	max. $\overline{\text{ACC}}$	ACC	p^*	max. $\overline{\text{ACC}}$	ACC	p^*
(a)	0.828	0.846	11.6	0.833	0.836	9.0
(b)	0.868	0.879	11.8	0.848	0.856	11.2
(c)	0.849	0.854	11.8	0.815	0.824	10.6

Table 6.5: *Classification accuracy using the genetic algorithm and* $E_j^{seg} = d_j$. Similar to the SDA (cf. Table 6.4), the feature reduction realised by means of the genetic algorithm has considerably improved the accuracies of all systems in comparison to the unreduced feature sets (cf. Table 6.3). The size of the feature sets can be diminished noticeably until a statistically significant reduction of the accuracies occurs. The respective feature numbers are denoted by p^*.

The results given in Table 6.5 show that the proposed genetic algorithm is an appropriate means to reduce the size of the feature set; the resulting systems are comparable to the SDA (cf. Table 6.4) and outperform the systems which do not employ feature reduction (cf. Table 6.3). The corresponding confusion matrix of an exemplary system is given by Table C.7 in Appendix C.4. So, this conclusion can be confirmed. However, the SDA seems to allow for a stronger reduction of the feature set (cf. Table 6.4 and Table 6.5).

As in the experiments with the unreduced feature sets, the classifiers using feature set (b) performed best. Feature set (a), which does not encompass region-dependent texture features, achieved only total accuracies, which are slightly worse. However, these differences are not statistically significant ($\alpha{=}0.05$). But they might indicate a benefit from employing region-dependent texture features rather than Zernike moments. The accuracies for classifiers based on feature set (c), which consists of both region-dependent texture features and Zernike moments, exhibit a tendency to be slightly lower than the results regarding feature set (b) as well. Although these differences are not statistically significant for $\alpha{=}0.05$, they could originate from difficulties of the genetic algorithm to select appropriate features from larger feature sets.[11]

The differences between the results of the SFAM and the SHAM are not statistically significant. Nevertheless, the SHAM is slightly surpassed by the SFAM. This might indicate that the SFAM is more suited to be employed in conjunction with the genetic algorithm. The only exception is feature set (a). Here, the mean and total accuracies are roughly equal, but the SHAM can be employed with a smaller feature set. On the whole, both classifiers are applicable to the task at hand and the principal results regarding the effects of the feature selection, which were observed using the SDA, could be confirmed.

A closer look at the selected features reveals differences to the SDA. In order to enable a better comparison, the following results refer to the SFAM using feature set (b), which was analysed regarding the SDA above. First of all, the number of chosen features varies between 7 and 15 with respect to the five runs of the genetic algorithm. Therefore, the variations between the corresponding feature sets have increased as well. However, three features, which were regarded as critical

[10] significance level: $\alpha{=}0.05$

[11] Feature set (c) consists of 122 individual features, whereas feature set (b) encompasses only 73 features.

by the SDA, are usually selected as well: the entropy of the complete cell, the first value of the pattern spectrum and the third value of the pattern spectrum. However, the selected features comprise all types of features available in feature set (b): region-dependent texture features, pattern spectra, fractal features and histogram-based features. So, the assumption that the chosen feature types appropriately reflect the different properties of the considered protein distribution patterns is substantiated.

6.5.4 Comparison of the SDA and the Genetic Algorithm

Both feature selection methods achieved excellent results. But there are differences, which must be taken into account. The SDA yields a ranking of the features that allows for the selection of a feature set of a given size. In contrast, the genetic algorithm not only enables a ranking, it further determines weights, which are multiplied with the input vectors. As a result, the classifiers obtained by means of both algorithms behave differently for given sizes of the feature set (see Figure 6.5).

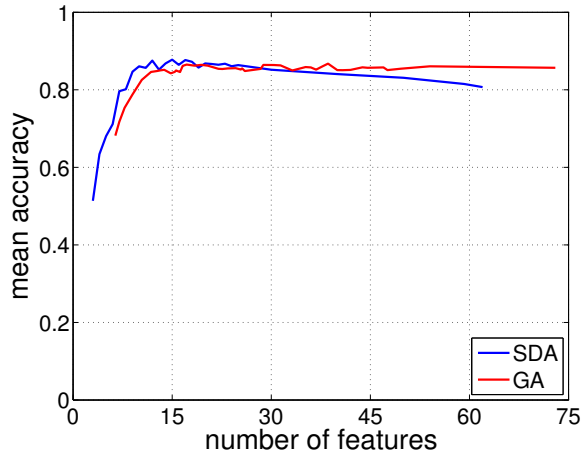


Figure 6.5: *Mean accuracies depending on the feature sets' sizes.* The SDA performs better for small feature sets, while the genetic algorithm achieves higher accuracies for larger sets. However, the maximum mean accuracies are roughly equal.

Although the overall accuracies are comparable, the systems based on the SDA seem to classify more correctly using very small feature sets. On the other hand, the genetic algorithm compensates for the decrease of the accuracies resulting from large feature sets.

The big advantage of the SDA is its speed. The complete processing of feature set (c), which comprises the maximum number of features (122), takes about 0.36s using an Mobile AMD Athlon XP processor operating at 557MHz. However, it must be considered that several assumptions are made. In particular the data are assumed to be multivariate normal distributed with a common covariance matrix [187, Chapter 67], which cannot be guaranteed. Nonetheless, the results summarised by Table 6.4 justify the SDA's application.

In contrast to the SDA, the genetic algorithm is rather computationally intensive. Even though computers using two Dual Core AMD Opteron processors (2GHz) were employed, which enable a parallel evaluation of four individuals, one run took at least 27 hours[12]. But depending on the applied system, the processing time rose up to 94 hours[13]. The fact that the genetic algorithm does not make assumptions on the data did not cause an improvement of the classification.

[12] SFAM, feature set (b)
[13] SHAM, feature set (a)

6.6 Learning of New Protein Locations

Besides the experiments discussed in Section 6.5, I analysed the ability of the proposed protein localisation technique to incorporate new protein locations after training. This is especially useful if new protein distribution patterns show up during an analysis performed by a biological expert. Such an expert could assign a new class label to the patterns, which enables their recognition and usage in further experiments. Therefore, the focus of my further work was on the recognition of unknown cell compartments and their learning in conjunction with an appropriate class label.

Using the simplified ARTMAP networks, as I propose in my thesis, an unknown location is characterised by a minimum activation $\tilde{z}^{F2}_{\min}(t)$ higher than a threshold τ. The corresponding feature vector has at least a distance of τ to the closest category. Here, the SFAM is based on the city block norm, whereas the SHAM employs the Euclidean distance.

However, τ is chosen in such a way as to maximise the mean accuracy (cf. Section 6.5). As a result, the vast majority of fluorescence images showing the protein locations used for training are considered as known and classified accordingly. So, if a new location resembles a known one, the corresponding feature vectors are likely to be regarded as known and classified incorrectly. In order to circumvent this problem, I introduced a second threshold τ_2 with respect to $\tilde{z}^{F2}_{\min}(t)$. It is smaller than τ and defines the minimum distance to the closest category that a pattern must have in order to be recognised as possibly unknown or rather as a possibly new protein location. In this case, an expert could be asked. τ_2 must be selected in such a way that a compromise between correctly classifying the known locations and detecting possibly unknown patterns is reached. In particular, the amount of required user interactions should be minimal. However, τ is still used for the classification. So, even if $\tilde{z}^{F2}_{\min}(t)$ for a specific feature vector is higher than τ_2, a suggestion for a likely protein location can be made unless it is higher than τ.

Section 6.6.1 introduces the evaluation procedures used to assess the classifiers' capabilities to detect and learn new protein locations. Additionally, the employed datasets are explained. Afterwards, the results are given in Section 6.6.2.

6.6.1 Evaluation Scheme

In order to simulate the occurrence of new protein location patterns during a biological experiment, I partitioned the available data according to the respective cell compartments: The patterns of eight protein locations were considered as known and utilised for training a classifier as explained in Section 6.5. The patterns showing the remaining two locations served as unknown distribution patterns for the evaluation. In principle, the application of a single cell compartment representing unknown patterns would have been possible as well. But the employed method is more realistic, since more than a single untrained location can be expected to exist. Unfortunately, there are 45 different possibilities of dividing ten classes into two groups of eight and two classes, respectively:

$$\binom{10}{8} = \binom{10}{2} = \frac{10!}{2!8!} = 45 \tag{6.18}$$

Hence, in total 45 combinations of the cell compartments have to be processed instead of ten. They are denoted by C_i (see Figure 6.6). However, an increment of the unknown locations from two to three would result in 110 different combinations. So, the chosen method constitutes a compromise between obtaining accurate results and reducing the computational load.

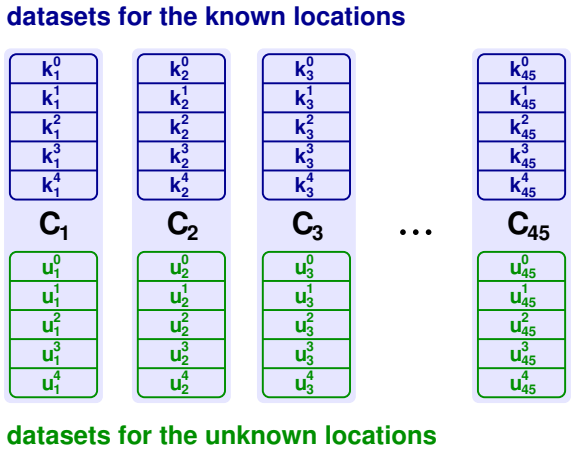


Figure 6.6: *Employed datasets.* For each combination C_i of known and unknown cell compartments ten datasets were created. They enable cross-validation with respect to each type of protein location.

Since the evaluation was performed using cross-validation, the datasets of each combination C_i of known and unknown locations were divided into five subsets. The individual datasets are referred to as k_i^j and u_i^j, respectively. The partitioning resulted in ten datasets regarding each combination and enables cross-validation with respect to the known as well as the new cell compartments.

The final evaluation procedure is rather complex, since for each of the 45 different combinations several training, validation and test runs are required. Therefore, its simplified pseudocode might be more illustrative:

```
forall combinations C_i {
  train a network using k_i^0,...,k_i^4
  for τ_2 ← 0 to τ_2^max {
    test which patterns of u_i^0,...,u_i^4 are unknown
    retrain with these patterns (one presentation)
    determine the classification results
  }
}
average the results over all C_i
determine the optimal value for τ_2
compute the corresponding mean accuracies
```

As mentioned above, cross-validation was performed with respect to both types of locations. But the corresponding loops have been omitted in the pseudocode in order to improve its comprehensibility. In particular, the classifiers were optimised using four-fold cross-validation: Three datasets were used for training and the fourth one for validation. The respective fifth dataset served as an independent test set for the determined parameter settings. This procedure was repeated for all possible groups of four datasets. On the whole, it is a simplification of the cross-validation scheme introduced in Section 6.5. This simplification became necessary, since additional evaluations are performed, which increase the computational load.

However, the cross-validation procedure yields five sets of parameters (ρ and τ), which need to be summarised so as to obtain a single network recognising the old locations. Since ρ as well as τ constitute some kind of distance measure, their final values are computed by averaging. Instead of utilising the settings resulting from cross-validation, these mean values are applied to testing.

So, the results are not completely independent from the respective training set. However, they constitute a good approximation of the real accuracy (cf. Sections 6.5).

The retraining is realised in a similar way: Four of five datasets are utilised. The remaining one is used as a independent test set. An optimisation of ρ and τ is not necessary here, as their values are already known. The whole process is repeated for every test set possible. So, it constitutes five-fold cross-validation.

Before retraining is performed, all input vectors of the new locations that are correctly recognised as unknown are determined. This constitutes the worst case. In practice, checking for unknown input vectors and retraining would occur in parallel. So, after a single input vector of a new class has been learned by the network, the corresponding location pattern is no longer unknown. As a result, some images of the corresponding cell compartment could already be classified correctly rather than being sorted out as unknown. In contrast, using the sequential procedure, all feature vectors contribute equally to the determination of τ_2, since the respective classifier is not modified depending on their precursors. Therefore, the sequential method has been favoured.

One important problem, which has to be dealt with, is the selection of an appropriate feature set. Unfortunately, it is impossible to chose features enabling the correct localisation of proteins in cell compartments or combinations of cell compartments that are not known in advance. Because of that, the features should cover all possible locations that might occur. A preceding feature selection step bears the risk of missing new protein location patterns. Therefore, I decided to employ an unreduced feature set (cf. Table 6.3). Here, feature set (b) has been proven to be most suited to the task at hand.[14] In order to limit the computational load of the experiments, I only employed $E_j^{\text{seg}}{=}d_j$ as similarity criterion for the automatic generation of additional patterns (see Section 6.5).

Besides the features themselves, the size of the feature set might affect the retraining process. Therefore, experiments based on reduced feature sets were conducted as well, although this kind of feature reduction would not be reasonable with respect to the detection and learning of new cell compartments.

6.6.2 Results

The assessment of individual classifiers is performed by means of several quantities. Firstly, the fraction of samples, which are classified as possibly unknown before retraining, is given. Here, a distinction between the old, known locations and the new, unknown cell compartments must be made. With respect to the unknown locations, this fraction, denoted by f^{u}, should be high. So, the new locations can be detected. On the other hand, images of already known classes should not be regarded as possibly unknown. The corresponding results are symbolised by f^{k}. Due to these considerations, the difference between f^{u} and f^{k} is maximised by varying τ_2. The value of τ_2 that leads to the maximum difference is referred to as τ_2^{opt} (see Figure 6.7).

In principle, alternative methods for determining an appropriate value of τ_2 would be possible. So, f^{k} could be diminished more strongly, if required. However, such a reduction entails a decline of f^{u} as well. Nevertheless, modifications of τ_2 might enable an adaptation of the proposed approach to a greater variety of tasks and users.

Besides the fraction of feature vectors classified as possibly unknown, the accuracy of the classifiers is critical. Hence, the mean accuracy is measured similar to Section 6.5. But here, the ac-

[14] Feature set (b) encompasses a total of 73 features.

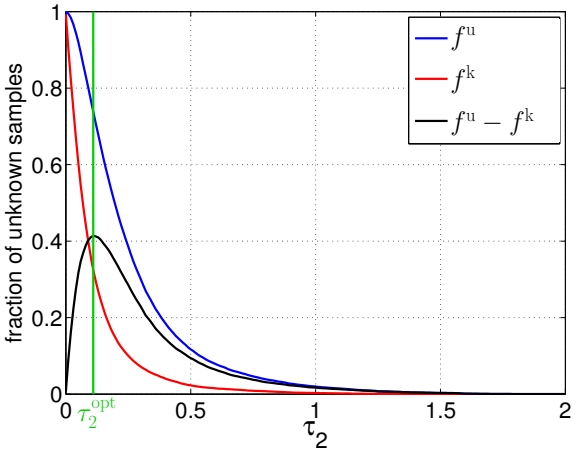


Figure 6.7: *Fractions of possibly unknown samples using the SFAM.* Depending on τ_2, different fractions of patterns from the known and from the new cell compartments are regarded as possibly unknown. Here, the value of τ_2, which maximises the difference between both fractions, is considered as optimal.

curacies regarding different subsets of the considered data are distinguished. $\overline{\text{ACC}}^{\text{u}}_{\text{test}}$ denotes the mean accuracy with respect to new protein locations after one epoch of retraining. Here, the test data were independent from the training sets. In contrast, $\overline{\text{ACC}}^{\text{u}}_{\text{incorrect}}$ symbolises the fraction of images showing new location patterns that were part of the respective training sets and incorrectly classified rather then being considered as possibly unknown. Hence, these images could not contribute to the retraining. Nevertheless, a certain fraction of them is classified correctly after other feature vectors of the corresponding class have been learned. Finally, $\overline{\text{ACC}}^{\text{k}}_{\text{before}}$ and $\overline{\text{ACC}}^{\text{k}}_{\text{after}}$ denote the mean accuracies of the known cell compartments before and after retraining, respectively.

Table 6.6 compares the results of both types of classifier under analysis – the SFAM and the SHAM. The given fractions of unknown patterns as well as the mean accuracies refer to the respective value of τ_2^{opt}.

classifier	τ_2^{opt}	f^{u}	f^{k}	$\overline{\text{ACC}}^{\text{u}}_{\text{test}}$	$\overline{\text{ACC}}^{\text{u}}_{\text{incorrect}}$	$\overline{\text{ACC}}^{\text{k}}_{\text{before}}$	$\overline{\text{ACC}}^{\text{k}}_{\text{after}}$
SFAM	0.11	0.739	0.325	0.719	0.583	0.848	0.829
SHAM	0.05	0.558	0.246	0.401	0.290	0.730	0.442

Table 6.6: *Results of the retraining.* Here, the fractions of unknown feature vectors (f^{u} and f^{k}) as well as the mean accuracies for several subsets of the data are given. All results refer to the value of τ_2^{opt} given in the second column. In terms of the mean accuracy, the SHAM is clearly outperformed by the SFAM.

In order to assess the results, different aspects have to be considered: Are the new locations correctly detected? Are the known locations regarded as known? Is an appropriate retraining possible? How does the mean accuracy of the old cell compartments develop? These questions are answered in the following.

Since a high fraction of new patterns being correctly detected as possibly unknown is a prerequisite to retrain the network, f^{u} should be high. But it has to be kept in mind that a compromise between f^{u} and f^{k} is necessary. Hence, the SFAM's value for f^{u} is still regarded sufficient. But the SHAM led to an inadequate result. On the other hand, the fraction f^{k} of feature vectors from known locations that are classified as possibly unknown should be as low as possible so as to reduce the amount of required user interactions. Here, the SHAM performed better. But the SHAM's difference between f^{u} and f^{k} is considerably lower. This implies that the separation of input vectors from known and unknown locations, respectively, is impaired.

The SFAM's $\overline{\mathrm{ACC}}^{\mathrm{u}}_{\mathrm{test}}$ indicates that the SFAM is able to incorporate new location patterns. Even difficult patterns, which closely resemble the old cell compartments (represented by $1-f^{\mathrm{u}}$), are classified with an acceptable mean accuracy ($\overline{\mathrm{ACC}}^{\mathrm{u}}_{\mathrm{incorrect}}$). Here, the choice of τ_2^{opt} has proven beneficial again, since it enables a good compromise between the level of classification accuracy regarding the new protein locations and the need for user interaction symbolised by f^{k} (cf. Figure 6.7 and Figure 6.8).

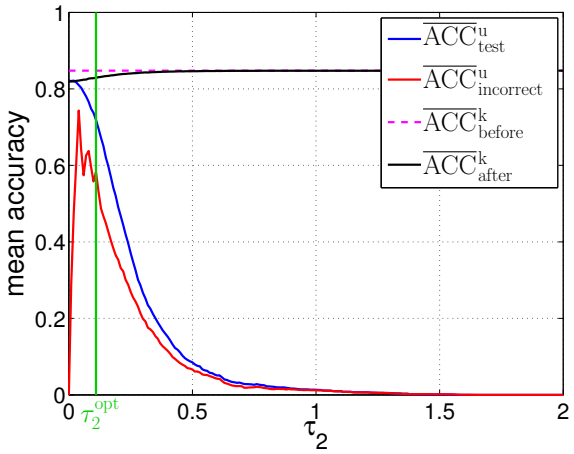


Figure 6.8: *Mean accuracies resulting from retraining the SFAM.* The mean accuracies for classifying the new cell compartments are strongly influenced by τ_2. Here, the proposed threshold τ_2^{opt} leads to good results. The mean accuracies concerning the a priori known protein locations are only slightly affected by τ_2. Moreover, τ_2^{opt} is considerably lower than τ, as τ usually exceeds one.

Unfortunately, the new protein locations' mean accuracies are noticeably lower than the mean accuracies with respect to the old, known locations. This is caused by the training procedure: The networks were trained with the old cell compartments until their weights did not change between two subsequent epochs. So, conflicts resulting from overlapping categories belonging to different classes are solved. Afterwards, the input vectors of the new locations were presented once. Therefore, the new locations are not as well integrated into the classifier as the old ones. This could be circumvented, if samples for the known cell compartments were presented during retraining. Then, there would be no difference between batch and on-line learning. Furthermore, the fraction of samples measured by f^{k} could be employed for training with samples of the old classes. So, it constitutes no drawback, if f^{k} does not equal zero. Finally, the retraining of the SFAM did not impair the classification accuracy of the previously known cell compartments. Thus, it can be concluded that the SFAM is applicable to tasks requiring the detection and incorporation of new protein locations.

In contrast to the SFAM, the results of the SHAM do not suffice: The new locations cannot be integrated in such a way that allows for their correct classification. Furthermore, the process of retraining considerably decreases the accuracy of the old classes. This is likely to be caused by the SHAM's category shape or the Euclidean distance applied to detect unknown feature vectors, since the SHAM's principal functioning is identical to the SFAM. Hence, the SFAM should be favoured.

In Section 6.5 and Section 5.4.5 was shown that the number of available features has a strong impact on the accuracy of the simplified ARTMAP networks. Hence, I assumed that it might affect the retraining process as well. Therefore, I arranged a reduced feature set. However, it is virtually impossible to select a set of features, which accounts for a number of classes that are not known in advance. Therefore, every kind of feature reduction potentially complicates the classification task. Even the complete feature sets might not be able to capture important information, although I

tailored it to a great number of cell compartments that might show up. However, a reduced feature set may reveal important information regarding the retraining process.

In order to determine a reduced feature set, I resorted to the systems described in Section 6.5. For comparison reasons, here, the classifiers using feature set (b) and the similarity criterion $E_j^{\text{seg}}{=}d_j$ were considered again. Unfortunately, all results were computed by means of cross-validation; that is, five runs using different classifiers were performed to determine mean values for the classification results. Consequently, five reduced feature sets resulted from each combination of a complete feature set, a classifier and a similarity criterion. So, a single reduced feature set needed to be compiled first. Here, I exploited the fact that the reduced feature sets for a specific parameter setting might be very similar, in particular, if the SDA is employed. The final features were therefore selected by a set union of the five sets yielded by the cross-validation runs. The size of these feature sets was chosen so as to maximise the mean accuracy (cf. Table 6.4). Due to the strong overlap of the five sets provided by the SDA, which comprise 15 features each, the resulting set encompasses only 22 features. As the SDA is independent of the chosen classifier, the SFAM and the SHAM employ the same feature set. The results of the retraining based on these features are given in Table 6.7

classifier	τ_2^{opt}	f^{u}	f^{k}	$\overline{\text{ACC}}^{\text{u}}_{\text{test}}$	$\overline{\text{ACC}}^{\text{u}}_{\text{incorrect}}$	$\overline{\text{ACC}}^{\text{k}}_{\text{before}}$	$\overline{\text{ACC}}^{\text{k}}_{\text{after}}$
SFAM, SDA	0.04	0.747	0.257	0.764	0.529	0.885	0.871
SHAM, SDA	0.00	0.783	0.366	0.705	0.195	0.831	0.809

Table 6.7: *Results of the retraining process using the SDA for feature reduction.* This table shows the fractions of unknown feature vectors (f^{u} and f^{k}) as well as the mean accuracies for several subsets of the data. All results refer to the value of τ_2^{opt} given in the second column. In terms of the mean accuracy, the SHAM is still outperformed by the SFAM. However, the feature selection seems to have a positive effect on the retraining capabilities of both networks.

A comparison of Table 6.6 and Table 6.7 reveals that the feature reduction improves the retraining capabilities. Here, the detection of unknown features and the mean accuracies could benefit. Now, the results of the SHAM are lying in a range which would enable its application as well. However, it must be kept in mind that the features were chosen using all ten cell compartments. In the case that a completely new protein location shows up, they might fail. Nevertheless, Table 6.7 indicates that the applied sets of features should be small. But in order to determine these features, expert knowledge is absolutely required.

6.7 Summary

This chapter has addressed the problem of localising tagged proteins in fluorescence micrographs that show protein distribution patterns of living Sf9 cells. The investigation of such location patterns enables the analysis of dynamic cellular processes. Here, the protein distribution patterns are described by sets of meaningful features. Based on these features, a classification is performed. So each protein location receives a specific class label, which corresponds to the observable cell compartments.

At first, a set of ten fixed classes was considered. They capture the majority of cell compartments, which can be distinguished without any additional co-localisation experiments (cf. Section 6.1). Here, the mean accuracy reached values of up to 88% (cf. Table 6.4). Hence, it can be concluded that proteins can be correctly recognised in the considered cell compartments. Further-

more, the basic feature sets were reduced so as to increase the classification accuracy. So, the most important features could be identified.

Besides the recognition of a fixed set of protein locations, the detection of and the retraining with new, unknown classes were investigated. Here, the applied classifiers were chosen in such a way as to enable incremental learning. By means of the introduction of appropriate thresholds regarding the minimum activation $\tilde{z}^{F2}_{\min}(t)$, they became capable of recognising unknown cell compartments. So, in addition to employing a fixed set of classes, potential users can semi-automatically incorporate new data if desired. In contrast to known approaches to the automatic identification of new protein locations, which are solely based on spatial relationships of the data in the feature space (cf. Section 6.1), the proposed approach guarantees that the generated classes have a biological meaning. Nonetheless, the inherent properties of the data are exploited so as to give suggestions. Furthermore, the consistency with the already known classes is ensured, which would barely be possible otherwise.

Since the protein localisation approach is intended to be employed for automated high-throughput experiments, its cooperation with the cell recognition method introduced in Chapter 5 is crucial. In order to facilitate this cooperation and to increase the amount of training data, automatically segmented cells were included in the training of the classifiers for the protein localisation. These cells had been obtained by means of the segmentation procedure described in Section 5.3. The results regarding these segments usually slightly surpass the accuracies with respect to the manually extracted cells, which is likely to originate from the higher amount of corresponding training samples (see Section 6.5). As a result, the developed cell recognition technique can be applied in conjunction with the proposed protein localisation approach without causing any problems.

7 Discussion and Outlook

Proteins are macromolecules, which play a major role for our existence. But although the DNA of several organisms, including humans, has been sequenced, our knowledge about proteins is rather limited; for example, only little is known about their functions, locations and interaction partners based solely on what has been learned from hereditary information. A living cell is a highly dynamic system that is influenced by a great variety of internal and external factors. These dynamic processes must be taken into account in order to characterise proteins. The knowledge of these processes and the involved proteins could be exploited to develop new remedies or innovative therapies, for example.

Location proteomics constitutes a very promising approach to the analysis of proteins in their natural environment – the living cell. Based on the location of a protein, which depends on the respective cell's state, conclusions about its function and interactions can be drawn. The proteins under analysis are usually fluorescently labelled. Then, they can be observed by means of a fluorescence microscope. The distribution of the labelled proteins allows their locations to be derived, i.e., the cell compartments, in which the proteins are located.

The goal of my thesis consists in providing methods for the automatic subcellular localisation of proteins in living cells. Due to the very high number of existing proteins, my focus was on techniques enabling high-throughput processing. I tackled two major problems that have to be solved in order to enable such a localisation of proteins: Firstly, I developed an approach to the automatic recognition of live cells in microscope images (see Chapter 5). Here, two different cell types, namely Sf9 cells and S2R+ cells, were considered. Furthermore, two microscopy techniques were employed: bright-field microscopy and differential interference contrast microscopy. These microscopy techniques are beneficial, since they allow for a parallel acquisition of fluorescence micrographs, which are required for analysing the protein distribution within the recognised cells.

In order to visualise the proteins under analysis, they are rendered visible by means of a fluorescent protein, which is fused on to them. The analysis of the corresponding protein distribution patterns constitutes the second problem which had to be solved (see Chapter 6). Here, a technique was developed, which not only allows for the recognition of a fixed set of protein locations, but enables the detection and incorporation of new patterns as well. This constitutes a major advantage in comparison to known techniques, which usually distinguish between techniques for recognising known and for identifying new protein locations (cf. Section 6.1).

Both methods – the cell recognition and the protein localisation – were designed in such a way as to facilitate their cooperation. Furthermore, except from labelling unknown protein locations for retraining, which is an optional extension of the proposed approach, user interactions are not necessary. So, a fully automatic system was developed. Here, several crucial adaptations of methods, which are known from the literature, for example active contours, had to be performed. In addition, established techniques such as SVMs were substituted with alternative methods, namely simplified ARTMAP networks. So, the system became able to detect unknown samples and learn them if required. A more detailed discussion of the proposed approaches to cell recognition and protein localisation is given in Section 7.1 and in Section 7.2, respectively.

Finally, Section 7.3 addresses possible limitations of the introduced system. In particular, it is discussed how they could be circumvented in the future and which extensions of the proposed methods could be reasonable for achieving this goal.

7.1 Cell Recognition

The introduced cell recognition technique was designed in such a way that it can be employed in conjunction with fluorescence microscopy, which is required for the protein localisation. In particular, compatible microscopy techniques had to be selected as a basis. Here, bright-field microscopy and differential interference contrast microscopy were chosen (see Chapter 4). These techniques enable the parallel acquisition of fluorescence micrographs without changing the principal optical path or influencing the proteins under analysis. As a result, the observed images can be superimposed.

In bright-field and DIC images, the most important feature of unstained live cells is their membrane (see Figure 7.1). It separates them from other cells and their surroundings. Hence, this information is crucial to the cell recognition (see Chapter 5). Unfortunately, the cell membranes are not the only image structures that are visible, so the usage of well-known cell recognition techniques such as thresholding is prohibited. Moreover, due to the size of the considered cells in these images, the application of techniques which iteratively analyse image patches is impeded; the processing of image patches of the required size would be too computationally intensive and down-scaling the images would result in less distinct cell membranes. Therefore, a new cell recognition method was necessary.

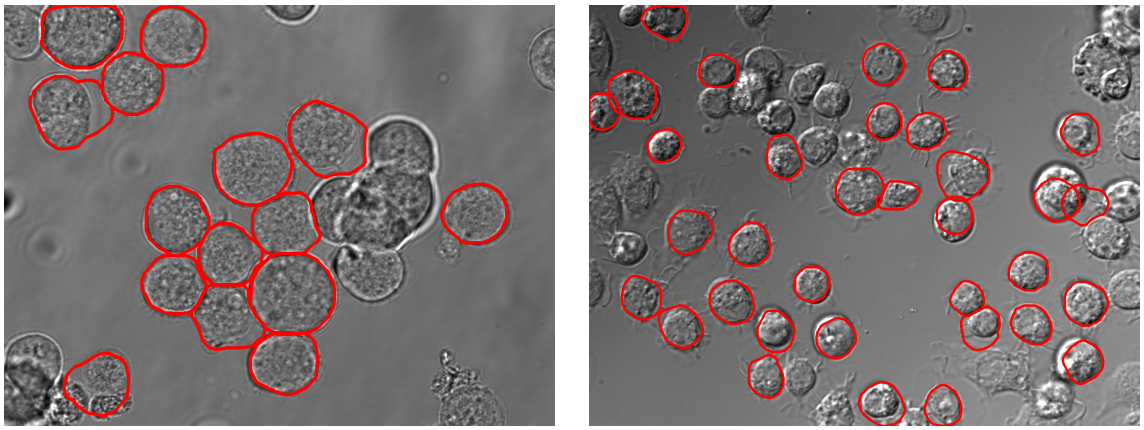


Figure 7.1: *Cell recognition results.* Both Sf9 cells (a) and S2R+ cells (b) can be correctly recognised, although they exhibit different characteristics. The snakes that correspond to found cells have been drawn as red contours. In contrast to Sf9 cells, which are recognised exclusively using bright-field images, the recognition of S2R+ cells requires DIC micrographs as well. Due to its improved contrast, here the DIC image has been employed to visualise the recognised S2R+ cells.

During the work on my thesis, I developed such a cell recognition procedure (see Chapter 5). It comprises three major steps: localisation, segmentation and classification (see Figure 5.2). This modular architecture allows for a substitution of the individual components, for example, if cells with different characteristics are to be analysed. So, although the proposed cell recognition technique was tailored to Sf9 cells, it could successfully be adapted to S2R+ cells. Figure 7.1 depicts

recognition results using both cell types. Here, the greedy snakes were employed to segment the cells.

The proposed cell recognition technique enables a correct recognition of up to 90% of the visible cells (cf. Table 5.5). Furthermore, the segmentation of these cells is sufficiently accurate to enable their application within the scope of the protein localisation.

7.2 Protein Localisation

Based on the recognised cells, the protein location patterns are analysed. Here, each pattern is associated with the cell compartments, in which the proteins are located. In order to reach this goal, I chose several sets of features, which seemed to adequately describe the protein distribution patterns observable in Sf9 cells. So, an efficient classification became possible. Here, a set of ten fixed protein locations was applied initially; i.e., the classifiers were trained using ten classes that they had to distinguish between subsequently. The corresponding mean accuracies reached values up to 88%, which indicates a very accurate classification of all ten cell compartments (cf. Table 6.4).

Since the proposed approaches to cell recognition and protein localisation are adapted to one another, their combined application is possible. In order to illustrate this ability, a corresponding image pair comprising a bright-field image and a fluorescence micrograph, which had been taken in parallel, was processed. The result is shown in Figure 7.2.

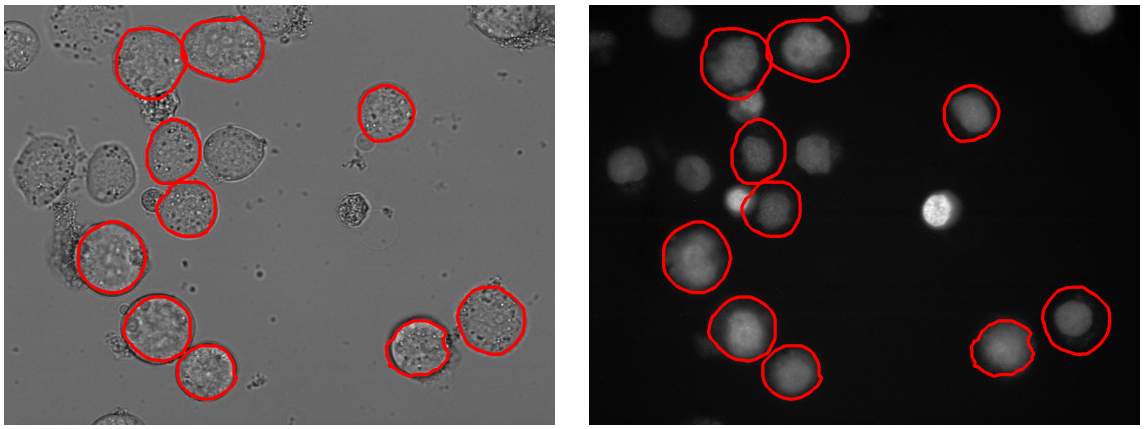


Figure 7.2: *Automatic protein localisation.* Based on a bright-field image (a), cells were recognised automatically (red contours). Afterwards, these cell segments were mapped to a corresponding fluorescence image showing tagged proteins (b), which are located in the cell nuclei. Finally, the classifier associated the protein distribution patterns of each cell segment with the class label representing the location 'nucleus'.

Although both considered classifiers – the SFAM and the SHAM – are in principle capable of on-line learning, their ability to detect and incorporate new protein locations after the basic training period was examined (see Section 6.6). Here, especially the SFAM achieved good results. By means of such a retraining, potential users are enabled to extend the system depending on their needs. This constitutes a major advantage in comparison to established approaches (cf. Section 6.1).

7.3 Outlook

In my thesis, I have introduced a system that enables the automated analysis of protein locations in living cells. It distinguishes between ten different protein locations. This number is comparable with alternative automated approaches to protein localisation, which are known from literature (see Section 6.1). But from a biological point of view, a more detailed partitioning is desirable; for example, the team of Won-Ki Huh manually assigned yeast proteins to one of 22 distinct locations [96]. They refined a set of 12 basic categories by means of co-localisation experiments with proteins, the localisations of which are known. In my opinion, the integration of such co-localisation information in an automatic protein localisation approach could enable a significant increment of the number of regarded protein locations as well. As an alternative, a small set of selected fluorescence dyes might be applicable.

However, when selecting additional tagged proteins or dyes, the goal of automatic processing must be kept in mind. In particular, the emission and excitation spectra of different stains should not overlap. Today, about four distinct fluorescent proteins can be applied in parallel without considerably influencing each other (cf. Section 3.5.1). Here, the effort for the development of a procedure that recognises cells in bright-field or DIC images pays off. If additional dyes were employed for cell recognition, the number of observable co-localisations would have been diminished. However, since one of the available fluorescent proteins is required to tag the proteins under analysis, only the three remaining fluorescent proteins are available for co-localisation experiments. The resulting three auxiliary location patterns must therefore be chosen in such a way as to facilitate the refinement of a large number of protein distribution patterns.

Additionally, extensions of the introduced cell recognition procedure are imaginable; for example, the image pairs used to recognise S2R+ cells (cf. Section 5.6) could be combined to single multi-channel images. The resulting more comprehensive view of the regarded cells might have a positive influence on the recognition. However, several operators would have to be substituted in order to allow for a processing of multi-channel images.

In principle, there exists no limitation of the proposed cell recognition technique, which restricts its field of application to insect cells. Alternative cell types could be employed as well. Nonetheless, the relevant information, for instance about pixels showing the cells' boundaries, must be available. Here, knowledge about the walls of plant cells could be applied, for example. By virtue of the modular architecture of the proposed cell recognition approach, such modifications allow for its application to a greater variety of biological problems.

Finally, the computational speed of both principal techniques – the cell recognition and the protein localisation – is an important property, since high-throughput experiments are intended. Although both techniques were designed in such a way that the computational load is kept low, further optimisations are possible and reasonable. With respect to the cell recognition, image pyramids could be applied. So, the time for processing a single bright-field image could be reduced considerably. Regarding the protein localisation approach, more research into efficiently computable features would be beneficial. As the comparison of the basic, unreduced feature sets (a), (b) and (c) has demonstrated, such a type of analysis can decrease the computational load remarkably (cf. Section 6.5). Here, it was shown that the computation of feature set (b) takes about half of the time required for computing feature sets (a) or (c) without decreasing the level of classification accuracy. But it is very likely that further accelerations are possible.

A Algorithms

This Appendix is dedicated to various algorithms which were required in order that a recognition of Sf9 cells could be realised. Appendix A.1 introduces a method for the selection of suitable values for τ_{bg} and n_{bg}, which are utilised for the extension of the image background during the cell localisation (cf. Section 5.3.2). Furthermore, an efficient method for its computation is proposed. Then, Appendix A.2 addresses the selection of an optimal length for the linear structuring elements necessitated by the applied membrane detection technique (cf. Section 5.3.3). Finally, a method enabling the reparametrisation of snakes is discussed in Appendix A.3 (cf. Section 5.3.7).

A.1 Extension of the Image Background

As discussed in Section 5.3.2, an extension of the image background after performing the separation of image foreground and background is beneficial. This is achieved utilising a geodesic erosion which is constrained by pixels with gradient magnitudes higher than a threshold τ_{bg} and a chosen number of iterations n_{bg}. Both τ_{bg} and n_{bg} can be determined automatically (see Appendix A.1.1). Furthermore, as the iterative computation of a geodesic erosion is a time-consuming operation, Appendix A.1.2 introduces an algorithm, which only considers relevant pixels. By that, the computation time is decreased significantly.

A.1.1 Automatic Determination of Required Parameters

τ_{bg} and n_{bg} are computed by means of a simple procedure using manually extracted cells. Here, two image regions are regarded for each cell mask: the mask image itself and a tube of 10 pixels around it. In principle, the background should be as close as possible to the cells. Hence, a high number of background pixels in the cell boundaries' proximity represented by the tube is desired. On the other hand, no or at least very few background pixels should occur in the area of a mask. Figure A.1 depicts the fractions of both regions which are covered by background pixels if τ_{bg} and n_{bg} are varied. Here, the results are averaged over 759 cell masks (cf. Section 5.3.7).

In order to reach a compromise between covering the tubes with background pixels and leaving the masks' areas untouched, I decided to choose τ_{bg} and n_{bg} in such a way that the number of background pixels is maximal and less than one percent of the masks' areas are considered as background. One result of the application of this approach is shown in Figure A.2.

Although the image background and foreground are separated more accurately, the impact on the cell segmentation is limited, as the growth of the snakes is stopped at the cell membrane which also specifies the mask image of the geodesic erosion. But the visible image quality is increased. Furthermore, it can be computed very efficiently. So I decided to use it nevertheless.

A.1.2 Efficient Computation

In order to increase the computational efficiency, the fact that the boundary between image foreground and background comprises a small number of pixels is exploited. These pixels are deter-

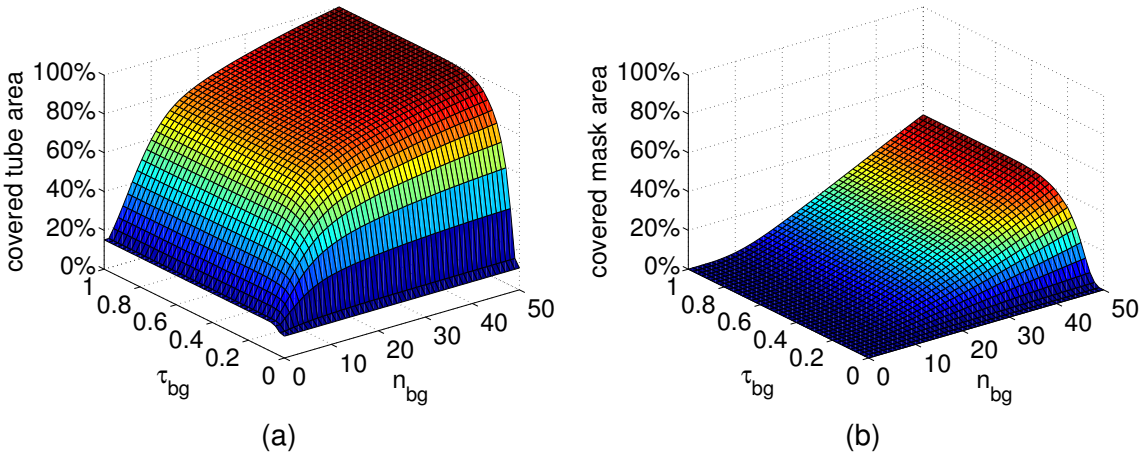


Figure A.1: *Covered area of the considered image regions depending on τ_{bg} and n_{bg}.* As the distance between the tube region and the initial image background is small, the number of background pixels within the tube is increasing fast if the background is extended (a). In contrast, the covered mask area is growing slowly (b).

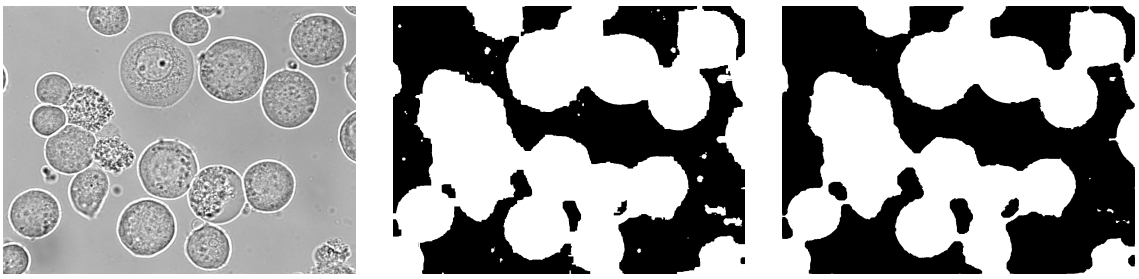


Figure A.2: *Extension of the image background by means of a geodesic erosion.* The application of the proposed geodesic erosion to an image (b) showing the foreground of a bright-field image (a) removes noise from the background region and leads to a more accurate approximation of the area covered by cells (c).

mined during the first iteration and put into a queue [204, Chapter 2]. In subsequent iterations only points from the queue are processed. The pseudocode of the algorithm is shown below. Here, $\mathcal{N}_4(\underline{x})$ denotes the neighbourhood of a point $\underline{x}$ using 4-connectivity and includes $\underline{x}$ itself.

```
# define NOT_IN_QUEUE      0
# define IN_QUEUE          1

input: background image I
       gradient magnitude image G
       threshold τ_bg
       number of iterations n_bg

output: eroded background image I

initialisations: queue, min_queue, point_queue

forall pixels x {                    // initialisation
  S(x) ← NOT_IN_QUEUE                // of the status
}                                    // image S
```

```
I' ← I                                  // copy image
forall pixels x {                       // fill queue
  if G(x) < τ_bg {
    min ← minimum(I'(N_4(x)))
    if min ≠ I(x) {
      I(x) ← min
      forall x' ∈ N_4(x) {
        if S(x') = NOT_IN_QUEUE {
          S(x') ← IN_QUEUE
          queue.enqueue(x')
        }
      }
    }
  }
}

for iter ← 1 to n_bg − 1 {
  while queue.empty() = false {         // process queue
    x ← queue.dequeue()
    S(x) ← NOT_IN_QUEUE
    if G(x) < τ_bg {
      min ← minimum(I(N_4(x)))
      if min ≠ I(x) {
        min_queue.enqueue(min){         // store for
        point_queue.enqueue(x)          // modification
      }
    }
  }

  while min_queue.empty() = false
        and point_queue.empty() = false {
    min ← min_queue.dequeue()
    x ← point_queue.dequeue()
    I(x) ← min                          // modify points
    if iter < n_bg − 1 {
      forall x' ∈ N_4(x) {              // store points
        if S(x') = NOT_IN_QUEUE {       // for the next
          S(x') ← IN_QUEUE              // iteration
          queue.enqueue(x')
        }
      }
    }
  }
}
```

For processing the main queue, two auxiliary queues are required. They are referred to as `min_queue` and `point_queue`, respectively. Without them, a separation of points from different iterations would be complicated.

In contrast to the definition of the geodesic dilation (see Section 5.3.1), no mask image is applied. In actual fact, the gradient magnitude image G is processed directly. But, a corresponding mask could be easily obtained by setting all values of $I(\underline{x})$ to zero if $G(\underline{x}) < \tau_{\mathrm{bg}}$.

The algorithm was evaluated based on 59 bright-field images (1344×1024 pixel) using the optimal parameters $\tau_{\mathrm{bg}} = 0.3$ and $n_{\mathrm{bg}} = 9$ (cf. Section 5.3.7) and an AMD Athlon 64 processor operating at 2GHz. In comparison to the original iterative approach, it enabled a reduction of the required computation time for the geodesic erosion of one bright-field image from 1.79s to 0.34s, on average.

A.2 Length of the Linear Structuring Elements

The basis for the automatic determination of the length l of the linear structuring elements required for the identification of possible cell membranes (cf. Section 5.3.3) is provided by n cell masks which were manually extracted by biological experts [231]. Besides the mask of a cell i itself, the points of a tube with a diameter of 5% of the mean cell diameter (9 pixels for Sf9 cells) that is centred at the masks' boundary are considered in order to detect the intensities of membrane pixels. The sets of the corresponding points p are denoted by $\mathcal{M}_i$ (mask) and $\mathcal{T}_i$ (tube), respectively. According to Equation A.3, then an optimal value for the length l of the line elements is computed by iterating over all reasonable values. Here, the intensities usually decline completely for lengths l that are smaller than Δ_{max}. Therefore, a maximum length of 99 pixels was considered.

$$I_l^{\mathcal{T}} = \sum_{i=1}^{n} \sum_{\forall p \in \mathcal{T}_i} I_l\big(x_p, y_p\big)^2 \tag{A.1}$$

$$I_l^{\mathcal{M}} = \sum_{i=1}^{n} \sum_{\forall p \in \mathcal{M}_i} I_l\big(x_p, y_p\big)^2 \Delta\big(x_p, y_p\big) \tag{A.2}$$

$$l_{\mathrm{opt}} = \arg\max_{\forall l} \left(\frac{I_l^{\mathcal{T}}}{\max\limits_{\forall l} I_l^{\mathcal{T}}} - \frac{I_l^{\mathcal{M}}}{\max\limits_{\forall l} I_l^{\mathcal{M}}} \right) \tag{A.3}$$

$I_l\big(x_p, y_p\big)$ constitutes the image generated by an algebraic opening with a structuring element of length l. The consideration of squared pixel values results in a reduced influence of small intensities that have less negative effects on the segmentation than high ones. Moreover, the points of the mask image are weighted by their minimal distance $\Delta\big(x_p, y_p\big)$ to the boundary. l_{opt} is optimal in a sense that it maximises the difference of the intensities (scaled to fit into the interval $[0, 1]$) within both examined image regions in order to enhance the contrast.

Figure A.3 visualises the determination of l_{opt} based on the relevant quantities. Here, 759 cells masks extracted from 59 bright-field images were regarded (cf. Section 5.3.7). In the case that l_{opt} is set to 31, the contrast between the cells' boundaries and interiors is maximised.

A.3 Insertion of New Snake Points

The segmentation consists in the extension of snakes starting from small regions within probable cells [231]. So, the distances between adjoining points are increased and resampling of the snake, i.e. the insertion of new points, is necessary. On the other hand, too high a number of points results

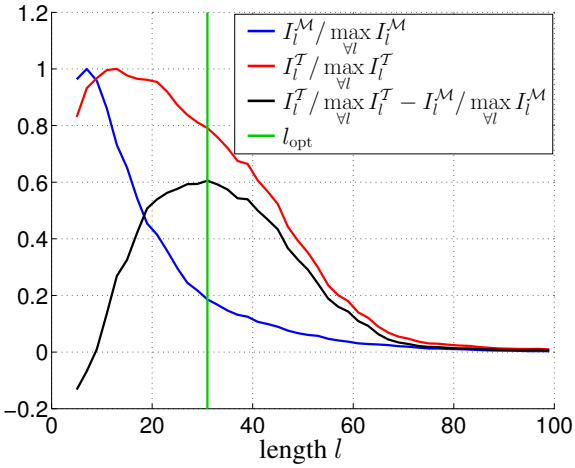


Figure A.3: *Determination of l_{opt}.* The optimal length l_{opt} is chosen in such a way as to maximise the intensity difference between the cells' interiors and boundaries.

in an increased computational effort. Thus, some kind of compromise has to be reached. Since the utilised Sf9 cells are almost elliptically shaped, an ellipse approximation of the current snake is performed [65]. This yields the lengths of the semiminor axis b and of the semimajor axis a as well as the centre C. On the basis of these values, the approximation error ϵ occurring if the ellipse is approximated by a line segment of length λ is computed (see Figure A.4).

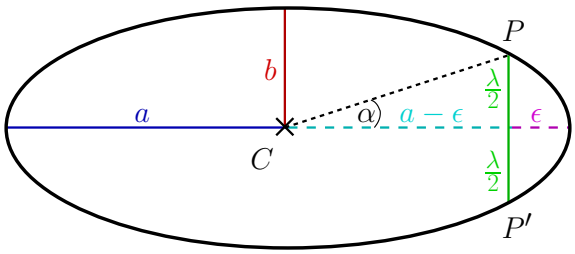


Figure A.4: *Approximation of an ellipse by line segments.* A line segment of length λ connecting the points P and P' causes an approximation error ϵ if it is equally divided by the semimajor axis. As the distance between the ellipse and its centre C is maximal there, ϵ is maximal as well. Thus, ϵ constitutes the worst case value.

An ellipse can be described by $x = a \cdot \cos\alpha$ and $y = b \cdot \sin\alpha$. Inserting the coordinates $x_P = a - \epsilon$ and $y_P = \frac{\lambda}{2}$ of point P and fusing the results leads to Equation A.4, which enables the determination of λ.

$$\lambda = 2b \cdot \sin\left(\arccos\frac{a-\epsilon}{a}\right) \tag{A.4}$$

Instead of computing the ellipse approximation after every iteration step of the snake algorithm (variable split length, VSL), it can be applied to the determination of a constant split length λ^* (CSL). For this purpose, manually extracted cells are approximated by ellipses and λ^* is set to the minimal value of λ. So, a correct approximation of all cells with an error less than ϵ can be guaranteed as well. The segmentation results of greedy snakes using VSL, CSL and no resampling are contrasted in Appendix C.1.

B Applied Features

Here, the features used for the discrimination between cell and non-cell segments are introduced. Depending on the cell type utilised, a specific feature set is employed. The Sf9-specific set is detailed in Appendix B.1. Appendix B.2 describes modifications of this set, which enable an improved recognition of S2R+ cells.

B.1 Recognition of Sf9 Cells

In order to realise a recognition of Sf9 cells in bright-field microscope images, a basic feature set comprising 111 features was chosen. These features allow for a separation of cell and non-cell segments. Three of these features reflect the shape of the obtained segments (see Appendix B.1.1). The remaining 108 features describe the histograms of several images' intensities in the considered segment-specific regions (see Section 5.4.1 and Appendix B.1.2). As only histograms are considered, these features are independent from the scale, the orientation and the position of the segments that are to be classified. They were chosen in such a way as to enable the incorporation of information on the position of the image background and potential cell membranes, which are crucial for a correct classification.

B.1.1 Shape features

Three features are employed in order to describe a segment's shape:

1. its area,
2. its perimeter
3. and its eccentricity.

So, the size of a segment as well as its deviation from a circular shape are measured and image regions exhibiting differing shapes can be sorted out.

While the first two features – the area and the perimeter – do not need any further explanation, the eccentricity necessitates some additional knowledge. It is a feature, which is based on *central moments* [103, Chapter 9]. These central moments enable a comprehensive description of shapes. A central moment μ_{pq} of order (p, q) is computed according to Equation B.1. Here $\mathcal{S}$ denotes the set of all points contained in a segment. $\bar{x}$ and $\bar{y}$ symbolise the coordinates of the corresponding centre of mass.

$$\mu_{pq} = \sum_{(x,y)\in\mathcal{S}} (x - \bar{x})^p (y - \bar{y})^q \tag{B.1}$$

Based on the notation of central moments, the eccentricity ε can be determined (see Equation B.2). It lies in the interval $[0, 1]$. A value of 0 represents a round object, whereas a value of 1

indicates a linear shape [109, Chapter 15].

$$\varepsilon = \frac{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}{(\mu_{20} + \mu_{02})^2} \tag{B.2}$$

B.1.2 Histogram-Based Features

Besides the three shape features, the grey-values of four types of images are examined. These images comprise the original bright-field micrograph, an image showing possible membrane pixels, its point-wise square and an image depicting the background. Based on these images, the following nine histogram-based features are analysed in four segment-dependent image regions (see Section 5.4.1):

1. the mean,
2. the variance,
3. the skewness (statistical moment of order 3),
4. the uniformity,
5. the smoothness,
6. the entropy,
7. the 5th percentile,
8. the 50th percentile (median)
9. and the 95th percentile.

The mean, the variance and the skewness constitute *statistical moments* of the grey-level distribution within the regarded image region [73, Chapter 11]. The mean μ represents the statistical moment μ_1 of order 1. Higher moments of order n are determined according to Equation B.3. The employed histogram encompasses L bins. The corresponding grey-levels are denoted by z_i and their frequencies by $p(z_i)$.

$$\mu_n = \sum_{i=0}^{L-1} (z_i - \mu)^n p(z_i) \tag{B.3}$$

The formulas for computing the uniformity u, the smoothness s and the entropy e can be derived from the histogram and the statistical moments (see equations B.4–B.6) [73, Chapter 11].

$$u = \sum_{i=0}^{L-1} p(z_i)^2 \tag{B.4}$$

$$s = 1 - \frac{1}{1 + \mu_2} \tag{B.5}$$

$$e = -\sum_{i=0}^{L-1} p(z_i) \log_2(z_i) \tag{B.6}$$

In order to compute the percentiles, the pixels of the considered image region are sorted with respect to their grey-levels, which results in a list with a length that is equivalent to the number of

inserted pixels. From this list, the grey-levels at 5%, 50% and 95% of the total length are selected, respectively. The 5th an the 95th percentiles constitute robust measures for the minimum and maximum intensities occurring. The 50th percentile represents the median.

Although, in principle, these nine features are applicable to the four segment-specific regions in all four considered images, two limitations were made: Firstly, the segment itself is only regarded in the original bright-field image. By this, an overall impression of the cell is enabled. This is intended to facilitate the rejection of objects that are not in focus. Based on the other image types, the utilised features computed for the segment itself hardly contribute to solving the classification problem at hand. Secondly, the percentiles were omitted regarding the background image, which depicts the areas that do not contain any cells. As this image contains only black and white pixels, the percentiles would have no reasonable meaning with respect to the task at hand. By virtue of these limitations, the number of histogram-based features is reduced from $4{\cdot}4{\cdot}9{=}144$ to $144{-}3{\cdot}9{-}3{\cdot}3{=}108$.

B.2 Recognition of S2R+-Cells

Similar to the features employed in order to recognise Sf9 cells, shape features as well as histogram-based features are utilised with respect to S2R+ cells. The shape features were transferred to the new problem. So they comprise a segment's area, its perimeter and its eccentricity. In contrast, the number of histogram-based features increased, since five rather than four segment-specific image regions are analysed (see Section 5.6.2). Regarding these regions, the following nine features, which are already known from the Sf9-specific feature set, are determined:

1. the mean,
2. the variance,
3. the skewness (statistical moment of order 3),
4. the uniformity,
5. the smoothness,
6. the entropy,
7. the 5th percentile,
8. the 50th percentile (median)
9. and the 95th percentile.

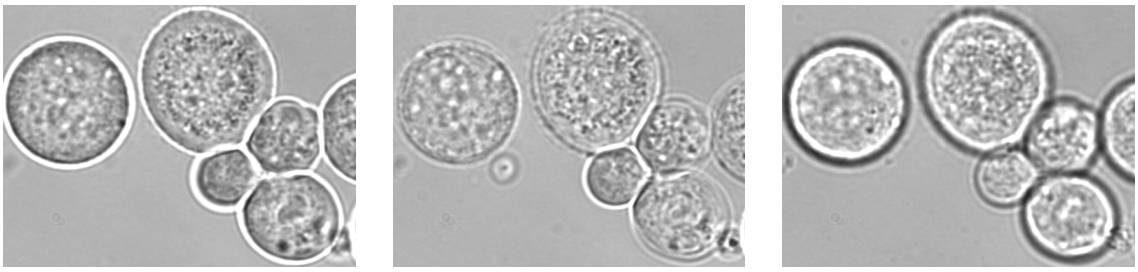
They are computed using the original bright-field micrograph, an image showing possible membrane pixels, its point-wise square and an image depicting the background. This yields a total of $4{\cdot}5{\cdot}9{=}180$ features. However, according to the discussion in Appendix B.1, several features do not contribute to solving the classification task at hand and can be omitted. Therefore, the segment itself is only regarded in the original bright-field image and the percentiles are not applied to the image that depicts the background. So only $180{-}3{\cdot}9{-}3{\cdot}4{=}141$ histogram-based features are applied.

C Further Analyses

In this appendix, the results of supplementary experiments are introduced. The investigations detailed in Appendix C.1 and Appendix C.2 were included in order to enable an optimisation of the microscope settings and allow for a better understanding of the segmentation results. In Appendix C.1, the segmentation results using different focal planes are discussed and Appendix C.2 addresses the deviations of manual cell segmentations performed by different people. Furthermore, Appendix C.3 introduces the confidence intervals for various classification tasks occurring in this thesis. So it becomes possible to assess the statistical significance of their differences. Eventually, Appendix C.4 gives the confusion matrices of several protein localisation approaches discussed in Section 6.5. These matrices enable conclusions regarding the recognition of specific protein location patterns to be drawn.

C.1 Usage of Different Focal Planes

In addition to the experiments discussed in Section 5.3.7, the segmentation method was evaluated with respect to different foci using a dataset containing 499 images of Sf9 cells which had been manually extracted from 45 images by biological experts [231]. This dataset comprised images of the same specimen at three manually adjusted focal planes (A, B and C) exhibiting the cell characteristics depicted in Figure C.1. All 499 manually extracted cells were automatically marked during the localisation step (see Section 5.3) and each cell mask was associated with the marker closest to its centre. Here the following settings were employed: Δ_{max}=198, τ_{bg}=0.3, n_{bg}=9 and l_{opt}=31. Their determination occurred according to Section 5.3. However, the maximum number of iterations was decreased from $\frac{1}{2}\Delta_{\text{max}}$ to $\frac{1}{4}\Delta_{\text{max}}$, since so the computation time is decreased significantly. Although the correct segmentation of very large cells cannot be guaranteed anymore, the results seem to be unaffected.



A B C

Figure C.1: *Cells at different focal planes.* The appearance of the examined cells varies if the focus is modified. Especially the characteristics of the cell membranes change.

In order to assess the segmentation, the manually extracted cell masks were compared with the corresponding automatically segmented cells by performing 15-fold cross-validation. The energy weights were chosen in such a way as to minimise the error term d_j (cf. Section 5.3.7) for all except one of the images of a focal plane. After computing the energy weights, the remaining

image was segmented in order to measure the test errors. μ_A, μ_B and μ_C denote the mean of these test errors over all images (see Table C.1). Additionally, the mean point number per snake $\bar{p}$ and the average processing time[1] $\bar{t}$ for a single image were determined. Here, an AMD Athlon 64 processor operating at 2GHz was employed.

method	ϵ	λ^*	μ_A	μ_B	μ_C	$\bar{p}$	$\bar{t}$
VSL	0.5	–	0.104	0.118	0.142	33.9	1.038s
	0.125	–	0.088	0.109	0.139	59.2	1.200s
CSL	0.5	18	0.094	0.130	0.143	45.2	0.802s
	0.125	9	0.102	0.116	0.141	89.6	0.980s
no resampling	–	–	0.109	0.123	0.146	23.4	0.708s

Table C.1: *Comparison of the segmentation if VSL, CSL, and no resampling are applied.* The dash denotes parameters that were not available. It is shown that the focal plane strongly influences the outcome of the segmentation, much more than the applied resampling method.

The results of all methods show that the choice of the focal plane has a considerable effect on the quality of the segmentation. In order to optimise the segmentation, focal plane A should be utilised for the acquisition of the bright-field images. Nevertheless, the results for plane B are still acceptable, especially if the proposed resampling methods with $\epsilon = 0.125$ are applied. At plane C, only errors of about 0.14 were achieved that originate from stronger intracellular intensity variations (see Figure C.1). Hence, plane C should not be used for cell segmentation, whereas plane A should be preferred.

Both reparametrisation methods attained smaller segmentation errors than the original approach, which does not perform resampling. Since CSL utilises a minimal value of the split length λ that is sufficient for all cells, it requires additional points in comparison to VSL. These unnecessary additional points seem to deteriorate the segmentation compared to VSL (e.g. for $\epsilon = 0.125$). The lowest errors were reached by VSL with $\epsilon = 0.125$, which required significantly more processing time than the other methods because of the determination of λ during the actual segmentation. So, if enough time is available, VSL should be employed. Otherwise, the original approach and CSL, especially with $\epsilon = 0.5$, is beneficial.

By virtue of the obtained results, the experiments conducted within the scope of this thesis employed VSL using $\epsilon = 0.125$ for resampling of the growing snakes. Furthermore, only images of cells taken at focal planes A and B were utilised. Here, plane B was included in order to improve the generalisation capabilities, although an exclusive usage of plane A could decrease the segmentation errors. But during the application of the proposed technique, the generalisation to similar cell appearances might be more important.

C.2 Deviations of Manual Segmentations

In order to assess the segmentation results, the manually extracted segments of 363 cells determined by five persons were compared pairwisely. These five persons segmented the cells using a common set guidelines. The deviations were measured based on the errors A_j and d_j (see Section 5.3.7). Here, a cell mask extracted by one person was considered as reference and a corresponding cell mask extracted by another person as an automatically determined segment. This

[1] excluding the time for the computation of the cell markers

procedure was repeated for all 363 available cells. Then the results were averaged over all pairwise combinations of human extractors. Combinations resulting in comparisons of persons with themselves were omitted, as the corresponding errors equal zero.

With respect to d_j, a mean error of 0.05 and a mean standard deviation of 0.025 were measured. Regarding A_j the mean error amounts to 0.083 and the standard deviation to 0.046. These variations of the manual segmentation indicate a lower error limit. In principle, automatic methods are not able to achieve better segmentations, as the boundary of the considered cells cannot be specified unambiguously.

C.3 Confidence Intervals of the Classification Results

In order to assess the statistical significance of the obtained classification results, I computed *confidence intervals*. Such confidence intervals specify a range of an estimated parameter, e.g. the mean or the variance of a probability distribution function, which encloses the true value with a given probability $(1-\alpha)$ called *confidence level* [189, Chapter 15]. Here, α denotes the *error probability*. It means that $100 \cdot (1-\alpha)\%$ of all possible confidence intervals estimated based on a fixed number of samples contain the true value of the parameter in question, while $100 \cdot \alpha\%$ do not.

Furthermore, it is assumed that the determined value is one of the $100 \cdot (1-\alpha)\%$ cases, which are enclosed by the computed interval. So it becomes possible to compare various results. Provided that the confidence intervals of two classification results obtained by means of different parameter settings do not overlap, their difference is statistically significant. Here, α is referred to as the *significance level* as well [189, Chapter 16]. In order to obtain significant results, I utilised α=0.05 within the scope of my thesis.

The computation of confidence intervals necessitates some knowledge about the probability distribution function of the underlying samples. With respect to classification tasks, the total accuracy ACC, which reflects the classifiers' performances (see Section 5.4.5), measures the number of correctly classified test samples; i.e., each sample can be considered as being either correctly or incorrectly classified. This can be modelled by means of a binomial distribution $\mathcal{B}(n,p)$, where n denotes the number of available samples and p the probability of correct classification results. The cell recognition rate can be analysed similarly.

However, it is more convenient to approximate the confidence interval by means of a standard normal distribution if n is sufficiently high [189, Chapter 15]. Assuming $n \geq 100$, the confidence interval of a binomially distributed parameter x is defined according to Equation C.1. Here, $\hat{x}$ denotes the estimated value of x and $z_{1-\frac{\alpha}{2}}$ the $100 \cdot \left(1-\frac{\alpha}{2}\right)$th percentile of the standard normal distribution, which can be taken from tables (e.g. Table C in [189]).

$$\left[\hat{x} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{x}(1-\hat{x})}{n}}, \hat{x} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{x}(1-\hat{x})}{n}}\right] \tag{C.1}$$

With respect to the task at hand, the accuracy ACC and the cell recognition rate equal $\hat{x}$. The confidence intervals belonging to the classification results of the cell recognition approach discussed in Chapter 5 are summarised in the following:

Table C.2 and Table C.3 are devoted to the recognition of Sf9 cells. They give the confidence intervals of the classification results shown in Table 5.3 and Table 5.4, respectively. As the error

probability α was set to 0.05, the confidence level equals 0.95. The evaluation occurred on the basis of the $3,878$ samples that had been determined manually.

		$E_j^{\text{seg}} = \max\left(A_j, d_j\right)$		$E_j^{\text{seg}} = d_j$	
classifier	feature reduction	ACC	confidence interval	ACC	confidence interval
SFAM	–	0.927	[0.919,0.935]	0.904	[0.895,0.913]
	CA	0.926	[0.918,0.934]	0.904	[0.895,0.913]
	CA & PCA	0.933	[0.925,0.941]	0.913	[0.904,0.922]
	CA & ICA	0.942	[0.935,0.950]	0.930	[0.922,0.938]
SHAM	–	0.896	[0.886,0.906]	0.876	[0.866,0.886]
	CA	0.906	[0.897,0.915]	0.873	[0.863,0.884]
	CA & PCA	0.911	[0.902,0.920]	0.884	[0.874,0.894]
	CA & ICA	0.916	[0.907,0.925]	0.908	[0.898,0.917]

Table C.2: *Confidence intervals of the simplified ARTMAPs' total accuracies (Sf9 cells).* In addition to the values of the accuracy shown in Table 5.3, here the confidence intervals are given. The were computed using an error probability of α=0.05. The red entries indicate values that are significantly better than the corresponding results of the same classifier using the complete feature set.

According to Table C.2, the approach combining the CA with the ICA enables a significant improvement of the accuracy in comparison to the complete feature set. Only for the SFAM in conjunction with E_j^{seg}=$\max\left(A_j, d_j\right)$ the difference is not statistically significant. But the respective confidence intervals barely overlap. So the results are not contradictory.

Table C.3 gives the confidence intervals corresponding to the SVCs' total accuracies shown in Table 5.4. In comparison to Table C.2, there is one major difference. The feature reduction does no longer cause an increment of the classification accuracy; it might even lead to a statistically significant reduction.

		$E_j^{\text{seg}} = \max\left(A_j, d_j\right)$		$E_j^{\text{seg}} = d_j$	
classifier	feature reduction	ACC	confidence interval	ACC	confidence interval
ν-SVC, linear	–	0.942	[0.935,0.950]	0.905	[0.896,0.914]
	CA	0.927	[0.919,0.935]	0.916	[0.907,0.925]
	CA & PCA	0.931	[0.923,0.939]	0.905	[0.896,0.914]
	CA & ICA	0.926	[0.918,0.934]	0.898	[0.889,0.908]
ν-SVC, RBF	–	0.942	[0.935,0.950]	0.919	[0.910,0.928]
	CA	0.933	[0.925,0.941]	0.909	[0.900,0.918]
	CA & PCA	0.930	[0.922,0.938]	0.899	[0.890,0.909]
	CA & ICA	0.931	[0.923,0.939]	0.905	[0.896,0.914]

Table C.3: *Confidence intervals of the SVCs' total accuracies (Sf9 cells).* In addition to the values of the accuracy shown in Table 5.4, the confidence intervals are given here. They were computed using an error probability of α=0.05. The blue entries indicate values that are significantly worse than the corresponding results of the same classifier using the complete feature set.

The confidence intervals regarding the classification of image segments depicting S2R+ cells are summarised in Table C.4. Here, only the SFAM was considered. For feature reduction, the combination of the CA and the ICA was chosen, since it yielded the best results concerning Sf9 cells. However, all experiments were conducted using bright-field as well as DIC images. So a total

of $1,450$ and $1,633$ manually determined samples, respectively, was available. Furthermore, an analysis of the influence of two similarity criteria for the automatic segment generation occurred.

feature set	image type	similarity criterion	ACC	confidence interval
Sf9-specific	bright-field	$d_j < 0.1$	0.850	[0.832,0.868]
		$d_j < 0.125$	0.844	[0.825,0.863]
	DIC	$d_j < 0.1$	0.870	[0.853,0.886]
		$d_j < 0.125$	0.873	[0.856,0.889]
S2R+-specific	bright-field	$d_j < 0.1$	0.878	[0.861,0.895]
		$d_j < 0.125$	0.863	[0.845,0.881]
	DIC	$d_j < 0.1$	0.893	[0.878,0.908]
		$d_j < 0.125$	0.890	[0.874,0.905]

Table C.4: *Confidence intervals of the total accuracies (S2R+ cells).* Here the confidence intervals of the total accuracies shown in Table 5.8 are given. The were computed using an error probability of α=0.05.

Using a significance level of α=0.05, the S2R+-specific feature set does not lead to any significant rise of the accuracies. However, the majority of the corresponding intervals overlap only slightly, which might indicate an improvement, nevertheless.

C.4 Confusion Matrices of the Protein Localisation Approaches

Besides the values of the mean accuracy and the total accuracy, the corresponding confusion matrices are an important means of analysing the quality of a protein localisation approach. They enable conclusions about similarities of different compartments to be drawn and can reveal deficiencies regarding the protein localisation approach; for example, the feature set might be unable to capture differences between certain protein distribution patterns. Therefore, some selected confusion matrices are given in the current appendix.

Table C.5 shows the confusion matrix of the neural network, which achieved the best classification results using the complete feature sets (cf. Table 6.3). This SFAM employed feature set (b) and the similarity criterion $E_j^{\text{seg}}=d_j$. Its total accuracy amounts to 83.3% and its mean accuracy to 81.8%.

The results summarised by Table C.5 indicate that the majority of the protein locations is found correctly. Some cell compartments, for example, the classes ‘cytoplasm including nucleus’ and ‘cytoplasm without nucleus’, are hard to distinguish. Here, it must be kept in mind that the underlying fluorescence micrographs themselves are ambiguous (cf. Figure 6.1). However, the results are very promising and prove that proteins can be localised in Sf9 cells using the proposed methods.

Table C.6 even shows a considerable improvement of the results from Table C.5. For comparison reasons, the SFAM using feature set (b) and $E_j^{\text{seg}}=d_j$ is regarded again. But here, a feature reduction using the SDA took place (cf. Table 6.4). The number of features was chosen in such a way as to maximise the mean accuracy. This resulted in a mean accuracy of $\overline{\text{ACC}}$=87.8% and a total accuracy of ACC=89.1%. In order to achieve these results, only 15 features were required. Nonetheless, the differentiation between the cell compartments was enhanced considerably.

The application of the SDA enables a better discrimination between very similar protein locations; for example, the classes ‘cytoplasm including nucleus’ and ‘cytoplasm without nucleus’ can

cell compartment	classification results (in percent)										
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(*)
cytoplasm incl. nucleus (a)	**88.9**	8.3	0.0	1.4	0.0	0.7	0.0	0.0	0.0	0.7	0.0
cytoplasm w/o nucleus (b)	28.6	**64.3**	1.8	0.0	3.6	0.0	0.0	0.0	0.0	1.8	0.0
endoplasmic reticulum (c)	0.0	0.0	**71.1**	2.1	4.2	21.1	1.4	0.0	0.0	0.0	0.0
lysosomes (d)	1.4	0.5	3.2	**82.0**	2.3	9.0	0.5	0.5	0.0	0.9	0.0
microtubules (e)	1.0	0.0	10.8	1.0	**84.3**	2.0	0.0	0.0	0.0	1.0	0.0
mitochondria (f)	0.4	0.0	11.2	6.7	2.2	**79.1**	0.0	0.0	0.0	0.4	0.0
nucleoli (g)	0.0	0.0	1.4	5.4	0.0	0.0	**79.7**	10.8	1.4	1.4	0.0
nucleus (h)	0.0	0.0	1.3	0.0	0.0	0.7	2.0	**96.0**	0.0	0.0	0.0
peroxisomes (i)	0.0	0.0	1.4	0.0	0.0	5.6	1.4	0.00	**87.3**	4.2	0.0
plasma membrane (j)	0.0	0.0	0.0	2.1	0.0	1.0	0.0	0.0	0.0	**96.9**	0.0

Table C.5: *Confusion matrix for the SFAM using the complete feature set (b) and* $E_j^{seg}=d_j$. Each row shows the classification results for a specific protein location. Therefore, a single entry denotes the fraction of images from a specific protein location (row), which were associated with a specific class label (column). In the case of a correct classification, the row index and the column index are equal. The column marked by (*) shows the fraction of protein distribution patterns which were considered as unknown.

be distinguished more accurately. So, the usage of the SDA with respect to a fixed set of considered protein locations is beneficial.

Similar to the SDA, the application of the genetic algorithm increases the classifiers' accuracies (see Table 6.5). This improvement is reflected by the confusion matrices as well. In order to enable a comparison, the SFAM using feature set (b) and $E_j^{\text{seg}}=d_j$ is analysed here. The feature set was selected so as to maximise the mean accuracy, which reached a value of $\overline{\text{ACC}}=86.8\%$. The corresponding total accuracy amounts to $\text{ACC}=87.9\%$. Table C.7 shows the confusion matrix.

However, with 38.6 features on average, the genetic algorithm requires a significantly higher number of features than the SDA to reach the maximum mean accuracy. This results from the fact that, in contrast to the SDA, the accuracies do not decline if large feature sets are applied (see Figure 6.5). Slightly lower results were achieved for considerably smaller feature sets as well. Using 17.2 features on average, the mean accuracy amounts to 86.5%, for example.

cell compartment	classification results (in percent) (a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(*)
cytoplasm incl. nucleus (a)	**97.2**	2.1	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
cytoplasm w/o nucleus (b)	17.9	**75.0**	1.8	0.0	3.6	0.0	0.0	0.0	0.0	1.8	0.0
endoplasmic reticulum (c)	0.0	0.0	**81.0**	2.8	7.0	7.0	0.7	1.4	0.0	0.0	0.0
lysosomes (d)	0.9	0.5	3.6	**85.6**	0.0	7.7	0.9	0.5	0.0	0.5	0.0
microtubules (e)	0.0	0.0	9.8	1.0	**87.3**	2.0	0.0	0.0	0.0	0.0	0.0
mitochondria (f)	0.0	0.0	4.1	4.5	0.0	**90.7**	0.0	0.0	0.4	0.4	0.0
nucleoli (g)	0.0	0.0	1.4	1.4	0.0	0.0	**86.5**	9.5	1.5	0.0	0.0
nucleus (h)	0.7	0.0	0.0	0.0	0.0	0.7	2.7	**96.0**	0.0	0.0	0.0
peroxisomes (i)	0.0	0.0	0.0	2.8	0.0	2.8	0.0	0.0	**88.7**	5.6	0.0
plasma membrane (j)	2.1	0.0	0.0	1.0	0.0	2.1	0.0	0.0	0.0	**94.9**	0.0

Table C.6: *Confusion matrix for the SFAM using $E_j^{seg}=d_j$ and 15 features selected from set (b) by means of the SDA.* Each row shows the classification results for a specific protein location. In comparison to Table C.5, which gives the results for the unreduced feature sets, the accuracies of the majority of the protein locations, e.g. the cytoplasm, have improved. Moreover, no input vectors are considered as unknown (last column).

cell compartment	classification results (in percent) (a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(*)
cytoplasm incl. nucleus (a)	**92.4**	5.6	0.0	1.4	0.0	0.0	0.0	0.7	0.0	0.0	0.0
cytoplasm w/o nucleus (b)	16.1	**76.8**	1.8	0.0	3.6	0.0	0.0	0.0	0.0	0.0	1.8
endoplasmic reticulum (c)	0.0	0.0	**80.3**	1.4	4.2	13.4	0.7	0.0	0.0	0.0	0.0
lysosomes (d)	0.0	0.5	3.2	**86.5**	0.9	7.7	1.4	0.0	0.0	0.0	0.0
microtubules (e)	0.0	1.0	7.8	2.0	**87.3**	1.0	0.0	0.0	0.0	1.0	0.0
mitochondria (f)	0.0	0.4	7.5	4.1	1.1	**86.6**	0.0	0.0	0.4	0.0	0.0
nucleoli (g)	0.0	0.0	2.7	4.1	0.0	1.4	**78.4**	10.8	1.4	1.4	0.0
nucleus (h)	0.0	0.0	0.0	0.0	0.0	0.7	1.3	**98.0**	0.0	0.0	0.0
peroxisomes (i)	0.0	0.0	0.0	1.4	0.0	4.2	0.0	0.0	**93.0**	1.4	0.0
plasma membrane (j)	0.0	0.0	0.0	3.1	0.0	2.1	0.0	0.0	1.0	**93.8**	0.0

Table C.7: *Confusion matrix for the SFAM using $E_j^{seg}=d_j$ and 38.6 features selected from set (b) by means of the genetic algorithm.* Each row shows the classification results for a specific cell compartment. As with the SDA, the accuracies of the majority of protein locations have risen compared to the systems using the unreduced feature sets. But now, some input vectors corresponding to cytoplasmic location patterns are classified as unknown. Nevertheless, this does not impair the overall results.

Bibliography

[1] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, et al. The genome sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000.

[2] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.

[3] H. Agaisse, L. S. Burrack, J. A. Philips, E. J. Rubin, N. Perrimon, and D. E. Higgins. Genome-wide RNAi screen for host factors required for intracellular bacterial infection. *Science*, 309:1248–1251, 2005.

[4] D. A. Agard, Y. Hiraoka, P. Shaw, and J. W. Sedat. Fluorescence microscopy in three dimensions. In *Methods in Cell Biology*, volume 30, pages 353–377. Academic Press, 1989.

[5] D. A. Agard, R. A. Steinberg, and R. M. Stroud. Quantitative analysis of electrophoretograms: A mathematical approach to super-resolution. *Analytical Biochemistry*, 111:257–268, 1981.

[6] L. G. Alexopoulos, G. R. Erickson, and F. Guilak. A method for quantifying cell size from differential interference contrast images: validation and application to osmotically stressed chondrocytes. *Journal of Microscopy*, 205:125–135, 2002.

[7] A. A. Amini, T. E. Weymouth, and R. C. Jain. Using dynamic programming for solving variational problems in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9):855–867, 1990.

[8] G. C. Anagnostopoulos and M. Georgiopoulos. Hypersphere ART and ARTMAP for unsupervised and supervised incremental learning. In *Proceedings of the International Joint Conference on Neural Networks*, volume 6, pages 59–64. 2000.

[9] G. C. Anagnostopoulos and M. Georgiopoulos. Ellipsoid ART and ARTMAP for incremental clustering and classification. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2, pages 1221–1226. 2001.

[10] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing – Partial Differential Equations and the Calculus of Variations*. Springer-Verlag, 2002.

[11] G. S. Baird, D. A. Zacharias, and R. Y. Tsien. Biochemistry, mutagenesis, and oligomerization of DsRed, a red fluorescent protein from coral. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):11984–11989, 2000.

[12] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.

[13] R. Beare. Regularized seeded region growing. In H. Talbot and R. Beare, editors, *Proceedings of the International Symposium on Mathematical Morphology (ISMM)*, pages 91–99. 2002.

[14] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[15] M. Bhasin, A. Garg, and G. P. S. Raghava. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, 21(10):2522–2524, 2005.

[16] M. V. Boland, M. K. Markey, and R. F. Murphy. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, 33:366–375, 1998.

[17] M. V. Boland and R. F. Murphy. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics*, 17(12):1213–1223, 2001.

[18] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 34(3):344–371, 1986.

[19] I. N. Bronstein and K. A. Semendjajew. *Taschenbuch der Mathematik*. B. G. Teubner, Nauka, 25th edition, 1991.

[20] G. Brumfiel. Who has designs on our students' minds? *Nature*, 434:1062–1065, 2005.

[21] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *Neural Networks*, 4:565–588, 1991.

[22] G. A. Carpenter, S. Grossberg, and J. H. Reynolds. ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, 4:565–588, 1991.

[23] G. A. Carpenter, S. Grossberg, and D. B. Rosen. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, 4:759–771, 1991.

[24] G. A. Carpenter, B. L. Milenova, and B. W. Noeske. Distributed ARTMAP: a neural network for fast distributed supervised learning. *Neural Networks*, 11:793–813, 1998.

[25] V. Caselles, F. Catté, T. Coll, and F. Dibos. A geometric model for active contours in image processing. *Numerische Mathematik*, 66:1–31, 1993.

[26] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[27] G. Cauwenberghs and T. Poggio. Incremental and decremental support vector machine learning. In *Advances in Neural Information Processing 13 (NIPS)*, pages 409–415. 2000.

[28] I. Chambers, J. Frampton, P. Goldfarb, N. Affara, W. McBain, and P. R. Harrison. The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the 'termination' codon, TGA. *The EMBO Journal*, 5(6):1221–1227, 1986.

[29] M. Chan, D. S. H. Tan, S.-H. Wong, and T.-S. Sim. A relevant in vitro eukaryotic live-cell system for the evaluation of plasmodial protein localisation. *Biochimie*, 88:1367–1375, 2006.

[30] C.-C. Chang and C.-J. Lin. *LIBSVM: a Library for Support Vector Machines*, June 2007. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[31] E. Chargaff. Structure and function of nucleic acids as cell constituents. *Federal Proceedings*, 10(3):654–659, 1951.

[32] A. Chebira, Y. Barbotin, C. Jackson, T. Merryman, G. Srinivasa, R. F. Murphy, and J. Kovačević. A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinformatics*, 8:210, 2007.

[33] S.-C. Chen, T. Zhao, G. J. Gordon, and R. F. Murphy. Automated image analysis of protein localization in budding yeast. *Bioinformatics*, 23:i66–i71, 2007.

[34] X. Chen and R. F. Murphy. Objective clustering of proteins based on subcellular location patterns. *Journal of Biomedicine and Biotechnology*, 2:87–95, 2005.

[35] X. Chen and R. F. Murphy. Interpretation of protein subcellular location patterns in 3D images across cell types and resolutions. In S. Hochreiter and R. Wagner, editors, *Proceedings of the International Conference on Bioinformatics Research and Development (BIRD)*, pages 328–342. Springer-Verlag, 2007.

[36] X. Chen, M. Velliste, and R. F. Murphy. Automated interpretation of subcellular patterns in fluorescence microscope images for location proteomics. *Cytometry*, 69A:631–640, 2006.

[37] X. Chen and C. Yu. Application of some valid methods in cell segmentation. In T. Zhang, B. Bhanu, and N. Shu, editors, *Proceedings of SPIE*, volume 4550, pages 340–344. 2001.

[38] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, 31(13):3497–3500, 2003.

[39] T.-J. Chin and D. Suter. Incremental kernel PCA for efficient non-linear feature extraction. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 3, pages 939–948. 2006.

[40] C. I. Christodoulou, C. S. Pattichis, E. Kyriacou, M. S. Pattichis, M. Pantziaris, and A. Nicolaides. Texture and morphological analysis of ultrasound images of the carotid plaque for the assessment of stroke. In L. Costaridou, editor, *Medical Image Analysis Methods*, pages 87–135. CRC Press, 2005.

[41] D. M. Chudakov, S. Lukyanov, and K. A. Lukyanov. Fluorescent proteins as a toolkit for in vivo imaging. *TRENDS in Biotechnology*, 23(12):605–613, 2005.

[42] A. Cichocki, J. Karhunen, W. Kasprzak, and R. Vigáro. Neural networks for blind separation with unknown number of sources. *Neurocomputing*, 24:55–93, 1999.

[43] L. D. Cohen. Note: On active contour models and balloons. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 53(2):211–218, 1991.

[44] C. Conrad, H. Erfle, P. Warnat, N. Daigle, T. Lörch, J. Ellenberg, R. Pepperkok, and R. Eils. Automatic identification of subcellular phenotypes on human cell arrays. *Genome Research*, 14:1130–1136, 2004.

[45] P. Cutler, G. Heald, I. R. White, and J. Ruan. A novel approach to spot detection for two-dimensional gel electrophoresis images using pixel value collection. *Proteomics*, 3:392–401, 2003.

[46] A. Danckaert, E. Gonzalez-Couto, L. Bollondi, N. Thompson, and B. Hayes. Automated recognition of intracellular organelles in confocal microscope images. *Traffic*, 3:66–73, 2002.

[47] H. Daniëlle, K. Cheung, H. E. Roossien, G. J. M. Pruijn, and J. M. H. Raats. A novel subtractive antibody phage display method to discover disease markers. *Molecular & Cellular Proteomics*, 5:245–255, 2006.

[48] C. Darwin. *The Origin of Species by Means of Natural Selection*. John Murray, sixth edition, 1926. Reprint of the edition from 1872.

[49] L. de Broglie. *Recherches sur la Théorie des Quanta*. Masson et Compagnie, 1963. Reprint of the edition from 1924.

[50] C. O. de Solorzano, R. Malladi, S. A. Lelièvre, and S. J. Lockett. Segmentation of nuclei and cells using membrane related protein markers. *Journal of Microscopy*, 201(3):404–415, 2001.

[51] O. Debeir, P. V. Ham, R. Kiss, and C. Decaestecker. Tracking of migrating cells under phase-contrast video microscopy with combined mean-shift processes. *IEEE Transactions on Medical Imaging*, 24(6):697–711, 2005.

[52] T. J. Deerinck, M. Martone, and M. H. Ellisman. Preparative methods for transmission electron microscopy. In D. L. Spector and R. D. Goldman, editors, *Basic Methods in Microscopy*, pages 303–306. Cold Spring Harbor Laboratory Press, 2006.

[53] C. P. Diehl and G. Cauwenberghs. SVM incremental learning, adaptation and optimization. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, volume 4, pages 2685–2690. 2003.

[54] M. Doverskog, J. Ljunggren, L. Öhman, and L. Häggström. Physiology of cultured animal cells. *Journal of Biotechnology*, 59:103–115, 1997.

[55] A. C. Doyle. *The Casebook of Sherlock Holmes*. BBC Consumer Publishing, 2005.

[56] G. Drewes and T. Bouwmeester. Global approaches to protein–protein interactions. *Current Opinion in Cell Biology*, 15(2):199–205, 2003.

[57] A. L. Edwards. *An Introduction to Linear Regression and Correlation*. W. H. Freeman and Company, 1976.

[58] M. H. Ellisman. Image production using transmission electron microscopy. In D. L. Spector and R. D. Goldman, editors, *Basic Methods in Microscopy*, pages 303–306. Cold Spring Harbor Laboratory Press, 2006.

[59] A. P. Engelbrecht. *Fundamentals of Computational Swarm Intelligence*. John Wiley & Sons, 2005.

[60] A. Fallert-Müller, editor. *Lexikon der Biochemie*. Spektrum Akademischer Verlag, first edition, 2000. Special edition for Weltbild, 2005.

[61] R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research*, 6:1889–1918, 2005.

[62] A. J. Feelders. Statistical concepts. In M. Berthold and D. J. Hand, editors, *Intelligent Data Analysis – An Introduction*, pages 17–68. Springer, second edition, 2003.

[63] G. A. Fink. *Mustererkennung mit Markov-Modellen*. Leitfäden der Informatik. B. G. Teubner, 2003.

[64] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[65] A. W. Fitzgibbon, M. Pilu, and R. B. Fisher. Direct least square fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(50):476–480, 1999.

[66] D. B. Fogel. *Evolutionary Computation – Toward a New Philosophy of Machine Intelligence*. John Wiley & Sons, third edition, 2006.

[67] M. Fussenegger, A. Opelt, and A. Pinz. Object localization/segmentation using generic shape priors. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 4, pages 41–44. 2006.

[68] J. J. Gagnepain and C. Roques-Carmes. Fractal approach to two-dimensional and three-dimensional surface roughness. *Wear*, 109:119–126, 1986.

[69] F. Galton. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45:135–145, 1888.

[70] E. García Osuna, J. Hua, N. W. Bateman, T. Zhao, P. B. Berget, and R. F. Murphy. Large-scale automated analysis of location patterns in randomly tagged 3T3 cells. *Annals of Biomedical Engineering*, 35(6):1081–1087, 2007.

[71] J. L. Gardy and F. S. L. Brinkman. Methods for predicting bacterial protein subcellular localization. *Nature Reviews Microbiology*, 4:741–751, 2006.

[72] C. C. Gillispie, editor. *Dictionary of Scientific Biography*, volume 2. Charles Scribner's Sons, 1973.

[73] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Pearson Education, second edition, 2002.

[74] R. C. Gonzalez, R. E. Woods, and S. L. Eddins. *Digital Image Processing Using MatLab*. Pearson Education, 2004.

[75] M. F. A. Goosen. Insect cell culture engineering: An overview. In *Insect Cell Culture Engineering*, pages 1–16. Marcel Dekker, 1993.

[76] M. Grobe, H. Volk, C. Münzenmayer, and T. Wittenberg. Segmentierung von überlappenden Zellen in Fluoreszenz- und Durchlichtaufnahmen. In *Proceedings of the Workshop "Bildverarbeitung für die Medizin" (BVM)*, pages 201–205. 2003.

[77] L. A. Gross, G. S. Baird, R. C. Hoffman, K. K. Baldridge, and R. Y. Tsien. The structure of the chromophore within DsRed, a red fluorescent protein from coral. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):11990–11995, 2000.

[78] S. Grossberg. How does a brain build a cognitive code? *Psychological Review*, 87(1):1–51, 1980.

[79] R. Gräf, J. Rietdorf, and T. Zimmermann. Live cell spinning disk microscopy. In *Advances in Biochemical Engineering/Biotechnology*, volume 95, pages 57–75. Springer, 2005.

[80] R. W. Y. Habash. *Electromagnetic Fields and Radiation – Human Bioeffects and Safety*. Marcel Dekker, 2002.

[81] R. W. Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 26(2):147–160, 1950.

[82] D. J. Hand. *Discrimination and Classification*. John Wiley & Sons, 1981.

[83] B. Hao, W. Gong, T. K. Ferguson, C. M. James, J. A. Krzycki, and M. K. Chan. A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science*, 296:1462–1466, 2002.

[84] D. Haussler, A. Krogh, I. S. Mian, and K. Sjolander. Protein modeling using hidden Markov models: Analysis of globins. In *Proceedings of the Hawaii International Conference on System Sciences*, pages 792–802. IEEE, Honolulu, 1993.

[85] D. D. Hegedus, T. A. Pfeifer, D. A. Theilmann, M. L. Kennard, R. Gabathuler, W. A. Jefferies, and T. A. Grigliatti. Differences in the expression and localization of human melanotransferrin in lepidopteran and dipteran insect cell lines. *Protein Expression and Purification*, 15:296–307, 1999.

[86] D. Hoffmann, H. Laitko, and S. Müller-Wille, editors. *Lexikon der bedeutenden Naturwissenschaftler*. Spektrum Akademischer Verlag, first edition, 2007. Special edition.

[87] C. Hoogland, K. Mostaguir, J.-C. Sanchez, D. F. Hochstrasser, and R. D. Appel. SWISS-2DPAGE, ten years later. *Proteomics*, 4(8):2352 – 2356, 2004.

[88] P. Horton, K.-J. Park, T. Obayashi, and K. Nakai. Protein subcellular localization prediction with WoLF PSORT. In *Proceedings of the Asian Pacific Bioinformatics Conference (APBC)*, pages 39–48. 2006.

[89] Y. Hu and R. F. Murphy. Automated interpretation of subcellular patterns from immunofluorescence microscopy. *Journal of Immunological Methods*, 290:93–105, 2004.

[90] Y.-H. Hu, D. Vanhecke, H. Lehrach, and M. Janitz. High-throughput subcellular protein localization using cell arrays. *Biochemical Society Transactions*, 33:1407–1408, 2005.

[91] J. Hua, O. N. Ayasli, W. W. Cohen, and R. F. Murphy. Identifying fluorescence microscope images in online journal articles using both image and text features. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1224–1227. 2007.

[92] F. Huang and J. Su. Moment-based shape priors for geometric active contours. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 2, pages 56–59. 2006.

[93] K. Huang and R. F. Murphy. Automated classification of subcellular patterns in multicell images without segmentation into single cells. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1139–1142. 2004.

[94] K. Huang and R. F. Murphy. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics*, 5:78, 2004.

[95] K. Huang, M. Velliste, and R. F. Murphy. Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. In *Proceedings of SPIE*, volume 4962, pages 307–318. 2003.

[96] W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425:686–691, 2003.

[97] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. D. Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. A. Sigrist. The PROSITE database. *Nucleic Acids Research*, 34:D227–D230, 2006.

[98] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[99] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[100] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945, 2004.

[101] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569–4574, 2001.

[102] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. Toward a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 97(3):1143–1147, 2000.

[103] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, 1989.

[104] R. I. Jennrich. Stepwise regression. In K. Enslein, A. Ralston, and H. S. Wilf, editors, *Statistical Methods for Digital Computers*, volume 3, pages 58–75. John Wiley & Sons, 1977.

[105] N. Jensen. *Etablierung und Charakterisierung von Sf9-Insektenzellen als Screening Organismus zur subzellulären Lokalisation von GFP-fusionierten Proteinen in der lebenden Zelle*. Diploma thesis, Faculty of Biology, Bielefeld University, Germany, 2005.

[106] X. C. Jin, S. H. Ong, and Jayasooriah. A practical method for estimating fractal dimension. *Pattern Recognition Letters*, 16:457–464, 1995.

[107] L. Jovine, S. Djordjevic, and D. Rhodes. The crystal structure of yeast phenylalanine tRNA at 2.0 Å resolution: Cleavage by Mg^{2+} in 15-year old crystals. *Journal of Molecular Biology*, 301:401–414, 2000.

[108] T. Junker. Charles Darwin und die Evolutionstheorien des 19. Jahrhunderts. In I. Jahn, editor, *Geschichte der Biologie*, pages 356–385. Spektrum Akademischer Verlag, third edition, 2002. Special edition for Nikol, 2004.

[109] B. Jähne. *Digitale Bildverarbeitung*. Springer-Verlag, 2002.

[110] N. A. Karp, J. L. Griffin, and K. S. Lilley. Application of partial least squares discriminant analysis to two-dimensional difference gel studies in expression proteomics. *Proteomics*, 5(1):81–90, 2005.

[111] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1987.

[112] P. M. Kasson, J. B. Huppa, M. M. Davis, and A. T. Brunger. A hybrid machine-learning approach for segmentation of protein localization data. *Bioinformatics*, 21(19):3778–3786, 2005.

[113] T. Kasuba. Simplified fuzzy ARTMAP. *AI Expert*, pages 18–25, November 1993.

[114] Z. Kawar, K. Karaveg, K. W. Moremen, and D. L. Jarvis. Insect cells encode a class II α-Mannosidase with unique properties. *Journal of Biological Chemistry*, 276(19):16335–16340, 2001.

[115] J. M. Keller and S. Chen. Texture description and segmentation through fractal geometry. *Computer Vision, Graphics, and Image Processing*, 45:150–166, 1989.

[116] J. D. Kelly and L. Davis. Hybridizing the genetic algorithm and the k nearest neighbors classification algorithm. In *Proceedings of the International Conference on Genetic Algorithms (ICGA)*, pages 377–383. 1991.

[117] A. Khotanzad and Y. H. Hong. Invariant image recognition by Zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):489–497, 1990.

[118] K. I. Kim, M. O. Franz, and B. Schölkopf. Iterative kernel principal component analysis for image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1351–1366, 2005.

[119] J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern Recognition*, 19(1):41–47, 1986.

[120] R. E. Kohler. *Lords of the Fly – Drosophila Genetics and the Experimental Life*. The University of Chicago Press, 1994.

[121] O. A. Koroleva, M. L. Tomlinson, D. Leader, P. Shaw, and J. H. Doonan. High-throughput protein localization in Arabidopsis using Agrobacterium-mediated transient expression of GFP-ORF fusions. *The Plant Journal*, 41(1):162–174, 2005.

[122] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 98–104. 1958.

[123] J. Kuzio and P. Faulkner. An overview of the molecular biology and applications of baculoviruses. In *Insect Cell Culture Engineering*, pages 17–50. Marcel Dekker, 1993.

[124] K.-C. Kwak and W. Pedrycz. Face recognition using an enhanced independent component analysis approach. *IEEE Transactions on Neural Networks*, 18(2):530–541, 2007.

[125] J.-B. Lamarck. *Philosophie Zoologigue*. H. R. Engelmann (J. Cramer) and Wheldon & Wesley, 1960. Reprint of the edition from 1809.

[126] F. Leymarie. Tracking and describing deformable objects using active contour models. Technical Report CIM-90-9, Computer Vision and Robotics Laboratory, McGill University, Montreal, Canada, 1990.

[127] U. Liebel, V. Starkuviene, H. Erfle, J. C. Simpson, A. Poustka, S. Wiemann, and R. Pepperkok. A microscope-based screening platform for large-scale functional protein analysis in intact cells. *FEBS Letters*, 554:394–398, 2003.

[128] K. Lindblad-Toh, C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe, M. Kamal, M. Clamp, J. L. Chang, E. J. Kulbokas III, M. C. Zody, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438:803–819, 2005.

[129] H. Liu and H. Motoda. Less is more. In H. Liu and H. Motoda, editors, *Feature Extraction, Construction and Selection – A Data Mining Perspective*, pages 3–12. Kluwer Academic Publishers, 1998.

[130] X. Long, W. L. Cleveland, and Y. L. Yao. Effective automatic recognition of cultured cells in bright field images using Fisher's linear discriminant preprocessing. *Image and Vision Computing*, 23:1203–1213, 2005.

[131] X. Long, W. L. Cleveland, and Y. L. Yao. Automatic detection of unstained viable cells in bright field images using a support vector machine with an improved training procedure. *Computers in Biology and Medicine*, 6(4):339–362, 2006.

[132] A. Macho, D. Decaudin, M. Castedo, T. Hirsch, S. A. Susin, N. Zamzami, and G. Kroemer. Chloromethyl-x-rosamine is an aldehyde-fixable potential-sensitive fluorochrome for the detection of early apoptosis. *Cytometry*, 25:333–340, 1996.

[133] M. T. Madigan, J. M. Martinko, and J. Parker. *Brock Biology of Microorganisms*. Pearson Education, tenth edition, 2003.

[134] N. Malpica, C. O. de Solórzano, J. J. Vaquero, A. Santos, I. Vallcorba, J. M.García-Sagredo, and F. del Pozo. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry*, 23:289–297, 1997.

[135] B. B. Mandelbrot. *Les Objets Fractals – Forme, hasard et dimension*. Flammarion, 1975.

[136] B. B. Mandelbrot. *The Fractal Geometry of Nature*. W. H. Freeman and Company, 1983.

[137] M. Mann and A. Pandey. Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *TRENDS in Biochemical Sciences*, 26(1):54–61, 2001.

[138] P. Maragos. Pattern spectrum and multiscale shape representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):701–716, 1989.

[139] G. Matheron. *Éléments pour une Théorie des Milieux Poreux*. Masson et Compagnie, 1967.

[140] M. V. Matz, A. F. Fradkov, Y. A. Labas, A. P. Savitsky, A. G. Zaraisky, M. L. Markelov, and S. A. Lukyanov. Fluorescent proteins from nonbioluminescent Anthozoa species. *Nature Biotechnlogy*, 17:969–973, 1999.

[141] H. Melderis. *Geheimnis der Gene – Die Geschichte ihrer Entschlüsselung*. Europäische Verlagsanstalt, 2001.

[142] H. Minkowski. Volumen und Oberfläche. *Mathematische Annalen*, 57:447–495, 1903.

[143] M. Minsky. Memoir on inventing the confocal scanning microscope. *Scanning*, 10:128–138, 1988.

[144] D. B. Murphy. *Fundamentals of Light Microscopy and Electronic Imaging*. Wiley-Liss, 2001.

[145] R. F. Murphy. Cytomics and location proteomics: Automated interpretation of subcellular patterns in fluorescence microscope images. *Cytometry*, 67A:1–3, 2005.

[146] R. F. Murphy. Systematic description of subcellular location for integration with proteomics databases and systems biology modeling. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1052–1055. 2007.

[147] R. F. Murphy, M. Velliste, and G. Porreca. Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. *Journal of VLSI Signal Processing*, 35:311–321, 2003.

[148] R. F. Murphy, M. Velliste, J. Yao, and G. Porreca. Searching online journals for fluorescence microscope images depicting protein subcellular location patterns. In *Proceedings of the IEEE International Symposium on Bioinformatics and Bioengineering (BIBE)*, pages 119–128. 2001.

[149] M. Nachtegael and E. E. Kerre. Classical and fuzzy approaches towards mathematical morphology. In M. Nachtegael and E. E. Kerre, editors, *Studies in Fuzziness and Soft Computing – Fuzzy Techniques in Image Processing*, pages 3–57. Physica-Verlag, 2000.

[150] J.-P. Nadal, E. Korutcheva, and F. Aires. Blind source separation in the presence of weak sources. *Neural Networks*, 13:589–596, 2000.

[151] T. W. Nattkemper, H. Wersing, H. Ritter, and W. Schubert. A neural network architecture for automatic segmentation of fluorescence micrographs. *Neurocomputing*, 48(4):357–367, 2002.

[152] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[153] G. Nomarski. Microintérferomètre différentiel à ondes polarisées. *Le journal de physique et le radium*, 16:9S–13S, 1954.

[154] S. Novianto, Y. Suzuki, and J. Maeda. Near optimum estimation of local fractal dimension for image segmentation. *Pattern Recognition Letters*, 24:365–374, 2003.

[155] E. Nägele, M. Vollmer, P. Hörth, and C. Vad. 2D-LC/MS techniques for the identification of proteins in highly complex mixtures. *Expert Review of Proteomics*, 1(1):37–46, 2004.

[156] Y. Oishi, S. Yunomura, Y. Kawahashi, N. Doi, H. Takashima, T. Baba, H. Mori, and H. Yanagawa. Escherichia coli proteome chips for detecting protein–protein interactions. *Proteomics*, 6:6433–6436, 2006.

[157] M. Ormö, A. B. Cubitt, K. Kallio, L. A. Gross, R. Y. Tsien, and S. J. Remington. Crystal structure of the Aequorea victoria green fluorescent protein. *Science*, 273:1392–1395, 1996.

[158] H. Paul, editor. *Lexikon der Optik*. Spektrum Akademischer Verlag, 1999.

[159] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[160] S. Peleg, J. Naor, R. Hartley, and D. Avnir. Multiple resolution texture analysis and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(4):518–523, 1984.

[161] A. P. Pentland. Fractal-based description of natural scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):661–674, 1984.

[162] P. Perner, S. Jänichen, and H. Perner. Case-based object recognition for airborne fungi recognition. *Artificial Intelligence in Medicine*, 36:137–157, 2006.

[163] A. Pierleoni, P. L. Martelli, P. Fariselli, and R. Casadio. BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, 22(14):e408–e416, 2006.

[164] M. Planck. Ueber das Gesetz der Energieverteilung im Normalspectrum. *Annalen der Physik*, 4(4):553–563, 1901.

[165] R. E. Plotnick, R. H. Gardner, W. W. Hargrove, K. Prestegaard, and M. Perlmutter. Lacunarity analysis: A general technique for the analysis of spatial patterns. *Physical Review E*, 53(5):5461–5468, 1996.

[166] T. Plötz. *Advanced Stochastic Protein Sequence Analysis*. Ph.D. thesis, Faculty of Technology, Bielefeld University, Germany, 2005.

[167] T. D. Pollard and W. C. Earnshaw. *Cell Biology*. Saunders, 2002.

[168] T. D. Pollard and W. C. Earnshaw. *Cell Biology*. Saunders, updated edition, 2004.

[169] E. M. Purcell, H. C. Torrey, and R. V. Pound. Resonance absorption by nuclear magnetic moments in a solid. *Physical Review*, 69:37–38, 1946.

[170] G. Qu, S. Hariri, and M. Yousif. A new dependency and correlation analysis for features. *IEEE Transactions on Knowledge and Data Engineering*, 17(9):1199–1207, 2005.

[171] A. T. Quiñones-Coello, L. N. Petrella, K. Ayers, A. Melillo, S. Mazzalupo, A. M. Hudson, S. Wang, C. Castiblanco, M. Buszczak, R. A. Hoskins, and L. Cooley. Exploring strategies for protein trapping in Drosophila. *Genetics*, 175(3):1089–1104, 2007.

[172] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, volume 77, pages 257–267. 1989.

[173] S. Raman, C. A. Maxwell, M. H. Barcellos-Hoff, and B. Parvin. Geometric approach to segmentation and protein localization in cell culture assays. *Journal of Microscopy*, 225:22–30, 2007.

[174] C. R. Rao. *Linear Statistical Inference and Its Applications*. John Wiley & Sons, 1965.

[175] N. Ray, S. T. Acton, and K. Ley. Tracking leukocytes in vivo with shape and size constrained active contours. *IEEE Transactions on Medical Imaging*, 21(10):1222–1235, 2002.

[176] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain. Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation*, 4(2):164–171, 2000.

[177] C. K. Reddy and F. B. Dazzo. Computer-assisted segmentation of bacteria in color micrographs. *Microscopy and Analysis*, 91:17–19, September 2004. European edition.

[178] S. Rey, M. Acab, J. L. Gardy, M. R. Laird, K. deFays, C. Lambert, and F. S. L. Brinkman. PSORTdb: a protein subcellular localization database for bacteria. *Nucleic Acids Research*, 33:D164–D168, 2005.

[179] P. G. Righetti and E. Boschetti. Sherlock Holmes and the proteome – a detective story. *The FEBS Journal*, 274(4):897–905, 2007.

[180] M. A. Rizzo, G. H. Springer, B. Granada, and D. W. Piston. An improved cyan fluorescent protein variant useful for FRET. *Nature Biotechnology*, 22(4):445–449, 2004.

[181] J. B. T. M. Roerdink and A. Meijster. The watershed transform: Definitions, algorithms and parallelization strategies. *Fundamenta Informaticae*, 41:187–228, 2000.

[182] K. F. A. Ross. *Phase Contrast and Interference Microscopy for Cell Biologists*. Edward Arnold, 1967.

[183] R.-A. Roszik. *Etablierung eines Mikroskopie-Systems zur automatisierten Lokalisierung von Zellstrukturen*. Diploma thesis, Faculty of Technology, Bielefeld University, Germany, 2005.

[184] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[185] W. R. Rypniewski, H. M. Holden, and I. Rayment. Structural consequences of reductive methylation of lysine residues in hen egg white lysozyme: An X-ray analysis at 1.8-Å resolution. *Biochemistry*, 32:9851–9858, 1993.

[186] N. Sarkar and B. B. Chaudhuri. An efficient approach to estimate fractal dimension of textural images. *Pattern Recognition*, 25(9):1035–1041, 1992.

[187] SAS Institute Inc. *SAS/STAT 9.1 User's Guide*. SAS Institute Inc., Cary, NC, 2004.

[188] D. A. Schiffmann, D. Dikovskaya, P. L. Appleton, I. P. Newton, D. A. Creager, C. Allan, I. S. Näthke, and I. G. Goldberg. Open Microscopy Environment and FindSpots: integrating image informatics with quantitative multidimensional image analysis. *BioTechniques*, 41:199–208, 2006.

[189] R. Schlittgen. *Einführung in die Statistik – Analyse und Modellierung von Daten*. R. Oldenburg Verlag, ninth edition, 2000.

[190] W. Schubert, M. Friedenberger, M. Bode, L. Philipsen, H. Ritter, and T. W. Nattkemper. Automatic recognition of muscle invasive T-lymphocytes expressing dipeptidyl-peptidase IV (CD26), and analysis of the associated cell surface phenotypes. *Journal of Theoretical Medicine*, 4:67–74, 2002.

[191] J. Schulz. Begründung und Entwicklung der Genetik nach der Entdeckung der Mendelschen Gesetze. In I. Jahn, editor, *Geschichte der Biologie*, pages 537–557. Spektrum Akademischer Verlag, third edition, 2002. Special edition for Nikol, 2004.

[192] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. Technical Report MSR-TR-99-87, Microsoft Research, 1999.

[193] B. Schölkopf, A. Smola, and K. R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[194] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimisation, and Beyond*. Adaptive Computation and Machine Learning. MIT Press, 2002.

[195] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[196] L. Scorrano, V. Petronilli, R. Colonna, F. D. Lisa, and P. Bernardi. Chloromethyltetramethylrosamine (Mitotracker Orange$^{\mathrm{TM}}$) induces the mitochondrial permeability transition and inhibits respiratory complex I. *The Journal of Biological Chemistry*, 274(35):24657–24663, 1999.

[197] J. A. Sethian. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge University Press, 1999.

[198] N. C. Shaner, P. A. Steinbach, and R. Y. Tsien. A guide to choosing fluorescent proteins. *Nature Methods*, 2(12):905–909, 2005.

[199] O. Shimomura, F. H. Johnson, and Y. Saiga. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea. *Journal of Cellular and Comparative Physiology*, 59:223–239, 1962.

[200] J.-B. Sibarita. Deconvolution microscopy. In *Advances in Biochemical Engineering/Biotechnology*, volume 95, pages 201–243. Springer, 2005.

[201] A. Sierra and A. Echeverría. Evolutionary discriminant analysis. *IEEE Transactions on Evolutionary Computation*, 10(1):81–92, 2006.

[202] R. Silipo. Neural networks. In M. Berthold and D. J. Hand, editors, *Intelligent Data Analysis – An Introduction*, pages 271–320. Springer, second edition, 2003.

[203] J. C. Simpson, R. Wellenreuther, A. Poustka, R. Pepperkok, and S. Wiemann. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO reports*, 1(3):287–292, 2000.

[204] S. S. Skiena. *The Algorithm Design Manual*. Springer-Verlag, 1998.

[205] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[206] P. Soille. *Morphological Image Analysis: Principles and Applications*. Springer-Verlag, 2003.

[207] P. Soille, E. J. Breen, and R. Jones. Recursive implementation of erosions and dilations along discrete lines at arbitrary angles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(5):562–667, 1996.

[208] M. Spitzer, S. Lorkowski, P. Cullen, A. Sczyrba, and G. Fuellen. IsoSVM – distinguishing isoforms and paralogs on the protein level. *BMC Bioinformatics*, 7:110, 2006.

[209] G. G. Stokes. On the change of refrangibility of light. *Philosophical Transactions of the Royal Society of London*, 142:463–562, 1852.

[210] J. V. Stone. *Independent Component Analysis – A Tutorial Introduction*. The MIT press, 2004.

[211] M. D. Summers and G. E. Smith. A manual of methods for baculovirus vectors and insect cell culture procedures. Texas Agricultural Experiment Station Bulletin No. 1555, 1987.

[212] Y. Suzuki. Self-organizing QRS-wave recognition in ECG using neural networks. *IEEE Transactions on Neural Networks*, 6(6):1469–1477, 1995.

[213] N. A. Syed, H. Liu, and K. K. Sung. Incremental learning with support vector machines. In *Proceedings of the Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence (IJCAI-99)*. 1999.

[214] M. Takeuchi and T. Ozawa. Methods for imaging and analyses of intracellular organelles using fluorescent and luminescent proteins. *Analytical Sciences*, 23(1):25–29, 2007.

[215] X. Tang, W. K. Steward, L. Vincent, H. Huang, M. Marra, S. M. Gallager, and C. S. Davis. Automatic plankton image recognition. *Artificial Intelligence Review*, 12:177–199, 1998.

[216] J. M. Tavaré, L. M. Fletcher, and G. I. Welsh. Using green fluorescent protein to study intracellular signalling. *Journal of Endocrinology*, 170:297–306, 2001.

[217] D. M. J. Tax and P. Juszczak. Kernel whitening for one-class classification. In S.-W. Lee and A. Verri, editors, *Proceedings of the International Workshop on Pattern Recognition with Support Vector Machines (SVM)*, pages 40–52. Springer, 2002.

[218] P. Taylor. Statistical methods. In M. Berthold and D. J. Hand, editors, *Intelligent Data Analysis – An Introduction*, pages 69–129. Springer, second edition, 2003.

[219] The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: A platform for investigating biology. *Science*, 282:2012–2018, 1998.

[220] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[221] The UniProt Consortium. The Universal Protein Resource. *Nucleic Acids Research*, 35:D193–D197, 2007.

[222] B. Thompson. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55(4):525–534, 1995.

[223] D. R. Thomsen, A. L. Meyer, and L. E. Post. Applications of insect cell gene expression in pharmaceutical research. In *Insect Cell Culture Engineering*, pages 105–138. Marcel Dekker, 1993.

[224] N. Tomiya, S. Narang, Y. C. Lee, and M. J. Betenbaugh. Comparing N-glycan processing in mammalian cell lines to native and engineered lepidopteran insect cell lines. *Glycoconjugate Journal*, 21:343–360, 2004.

[225] A. Tramontano. *The Ten Most Wanted Solutions in Protein Bioinformatics*. Chapman & Hall/CRC, 2005.

[226] M. Tscherepanow. *Vergleich von Antizipationsmechanismen hinsichtlich ihrer Fähigkeit zur Anpassung an veränderliche Systemaufgaben und Umweltbedingungen*. Diploma thesis, Faculty of Computer Science and Automation, Ilmenau Technical University, Germany, 2003.

[227] M. Tscherepanow and A. Heinze. Bildung bewertungsgesteuerter sensorischer Repräsentationen. Technical Report TR-NI-02-02, Schriftenreihe des FG Neuroinformatik der TU Ilmenau, 2002.

[228] M. Tscherepanow, N. Jensen, and F. Kummert. Recognition of unstained live Drosophila cells in microscope images. In *Proceedings of the International Machine Vision and Image Processing Conference (IMVIP)*, pages 169–176. IEEE, 2007.

[229] M. Tscherepanow and F. Kummert. Subcellular localisation of proteins in living cells using a genetic algorithm and an incremental neural network. In *Proceedings of the Workshop "Bildverarbeitung für die Medizin" (BVM)*, pages 11–15. Springer, 2007.

[230] M. Tscherepanow and A. Scheidig. FuzzyART-basierte Ansätze für sensomotorische Problemstellungen. In *Fortschrittberichte VDI Informatik/Kommunikationstechnik, VDI-Verlag: SOAVE'2004 – Selbstorganisation von adaptivem Verhalten*, pages 68–78. 2004.

[231] M. Tscherepanow, F. Zöllner, M. Hillebrand, and F. Kummert. Automatic segmentation of unstained living cells in bright-field microscope images. In *International Conference on Mass-Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry (MDA)*, volume 5108 of *LNAI*, pages 158–172. Springer, 2008.

[232] M. Tscherepanow, F. Zöllner, and F. Kummert. Aktive Konturen für die robuste Lokalisation von Zellen. In *Proceedings of the Workshop "Bildverarbeitung für die Medizin" (BVM)*, pages 375–379. Springer, 2005.

[233] M. Tscherepanow, F. Zöllner, and F. Kummert. Classification of segmented regions in brightfield microscope images. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 3, pages 972–975. IEEE, 2006.

[234] M. Tscherepanow, F. Zöllner, and F. Kummert. Segmentierung ungefärbter, lebender Zellen in Hellfeld-Mikroskopbildern. In *Proceedings of the Workshop "Bildverarbeitung für die Medizin" (BVM)*, pages 359–363. Springer, 2006.

[235] R. Y. Tsien. The green fluorescent protein. *Annual Review of Biochemistry*, 67:509–544, 1998.

[236] M. Twyman. *Principles of Proteomics*. BIOS Scientific Publishers, 2004.

[237] J. K. Udupa, V. R. LaBlanc, H. Schmidt, C. Imielinska, P. K. Saha, G. J. Grevera, Y. Zhuge, L. M. Currie, P. Molholt, and Y. Jin. Methodology for evaluating image-segmentation algorithms. In M. Sonka and J. M. Fitzpatrick, editors, *Proceedings of SPIE*, volume 4684, pages 266–277. 2002.

[238] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, et al. A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.

[239] E. R. Urbach, J. B. T. M. Roerdink, and M. H. F. Wilkinson. Connected rotation-invariant size-shape granulometries. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 1, pages 688–691. 2004.

[240] E. R. Urbach and M. H. F. Wilkinson. Shape-only granulometries and gray-scale shape filters. In *Proceedings of the International Symposium on Mathematical Morphology (ISMM)*, pages 306–314. 2002.

[241] M.-T. Vakil-Baghmisheh and N. Pavešić. A fast simplified fuzzy ARTMAP network. *Neural Processing Letters*, 17(3):273–316, 2003.

[242] M. van Herk. A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels. *Pattern Recognition Letters*, 13:517–521, 1992.

[243] V. N. Vapnik. Universal learning technology: Support vector machines. *NEC Journal of Advanced Technology*, 2(2):137–144, 2005.

[244] V. N. Vapnik and A. Y. Chervonenkis. On a class of perceptrons. *Automation and Remote Control, translated from "*Автоматика и Телемеханика*"*, 25(1):103–109, 1964.

[245] J. L. Vaughn, R. H. Goodwin, G. J. Tompkins, and P. McCawley. The establishment of two cell lines from the insect Spodoptera frugiperda (Lepidoptera; Noctuidae). *In Vitro*, 13(4):213–217, 1977.

[246] S. J. Verzi, G. L. Heileman, M. Georgiopoulos, and G. C. Anagnostopoulos. Universal approximation with fuzzy ART and fuzzy ARTMAP. In *Proceedings of the International Joint Conference on Neural Networks (IJCANN)*, volume 3, pages 1987–1992. 2003.

[247] B. Vigdor and B. Lerner. Accurate and fast off and online fuzzy ARTMAP-based image classification with application to genetic abnormality diagnosis. *IEEE Transactions on Neural Networks*, 17(5):1288–1300, 2006.

[248] L. Vincent. Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms. *IEEE Transactions on Image Processing*, 2(2):176–201, 1993.

[249] L. Vincent. Fast opening functions and morphological granulometries. In *Proceedings of SPIE*, volume 2300, pages 253–267. 1994.

[250] L. Vincent. Granulometries and opening trees. *Fundamenta Informaticae*, 41:57–90, 2000.

[251] R. F. Voss. Random fractals: Characterization and measurement. In *Scaling Phenomena in Disordered Systems*, pages 1–11. 1985.

[252] R. F. Walker, P. T. Jackway, and B. Lovell. Classification of cervical cell nuclei using morphological segmentation and textural feature extraction. In *Australian and New Zealand Conference on Intelligent Information Systems*, pages 297–301. 1994.

[253] G. Walsh. *Proteins – Biochemistry and Biotechnology*. John Wiley & Sons, 2002.

[254] J. Wang and D. C. Boisvert. Structural basis for GroEL-assisted protein folding from the crystal structure of (GroEL-KMgATP)$_{14}$ at 2.0 Å resolution. *Journal of Molecular Biology*, 327:843–855, 2003.

[255] Y.-L. Wang. Digital deconvolution of fluorescence images for biologists. In *Methods in Cell Biology*, volume 56, pages 305–315. Academic Press, 1998.

[256] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids – a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[257] N. Wei, J. You, K. Friehs, E. Flaschel, and T. W. Nattkemper. In situ dark field microscopy for on-line monitoring of yeast cultures. *Biotechnology Letters*, 29(3):373–378, 2007.

[258] E. R. Weibel. The significance of fractals for biology and medicine – an introduction and summary. In T. F. Nonnenmacher, G. A. Losa, and E. R. Weibel, editors, *Fractals in Biology and Medicine*, pages 2–6. 1993.

[259] Å. M. Wheelock and A. R. Buckpitt. Software-induced variance in two-dimensional gel electrophoresis image analysis. *Electrophoresis*, 26:4508–4520, 2005.

[260] D. J. Williams and M. Shah. A fast algorithm for active contours and curvature estimation. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 55(1):14–26, 1992.

[261] J. R. Williamson. Gaussian ARTMAP: a neural network for fast incremental learning of noisy multidimensional maps. *Neural Networks*, 9(5):881–897, 1996.

[262] W. D. Wilson, F. A. Tanious, H. J. Barton, L. Strekowski, and D. W. Boykin. Binding of 4',6-diamidino-2-phenylindole (DAPI) to GC and mixed sequences in DNA: Intercalation of a classical groove-binding molecule. *Journal of the American Chemical Society*, 111:5008–5010, 1989.

[263] N. Wiwatwattana, C. M. Landau, G. J. Cope, G. A. Harp, and A. Kumar. Organelle DB: an updated resource of eukaryotic protein localization and function. *Nucleic Acids Research*, 35:D810–D814, 2007.

[264] H. Wolff, K. Hadian, M. Ziegler, C. Weierich, S. Kramer-Hammerle, A. Kleinschmidt, V. Erfle, and R. Brack-Werner. Analysis of the influence of subcellular localization of the HIV Rev protein on Rev-dependent gene expression by multi-fluorescence live-cell imaging. *Experimental Cell Research*, 312:443–456, 2006.

[265] C.-M. Wu, Y.-C. Chen, and K.-S. Hsieh. Texture features for classification of ultrasonic liver images. *IEEE Transactions on Medical Imaging*, 11(2):141–152, 1992.

[266] K. Wu, D. Gauthier, and M. Levine. Live cell image segmentation. *IEEE Transactions on Biomedical Engineering*, 42(1):1–12, 1995.

[267] C. Wählby, P. Karlsson, S. Henriksson, C. Larsson, M. Nilsson, and E. Bengtsson. Finding cells, finding molecules, finding patterns. In P. Perner, editor, *Workshop on Mass-Data Analysis of Images and Signals (MDA)*, pages 15–24. IBaI CD-Report, 2006.

[268] C. Xu and J. L. Prince. Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing*, 7(3):359–369, 1998.

[269] S. Yanagawa, J.-S. Lee, and A. Ishimoto. Identification and characterization of a novel line of Drosophila Schneider S2 cells that respond to wingless signaling. *The Journal of Biological Chemistry*, 273(48):32353–32359, 1998.

[270] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems*, 13(2):44–49, 1998.

[271] D. Young, C. A. Glasbey, A. J. Gray, and N. J. Martin. Towards automatic cell identification in DIC microscopy. *Journal of Microscopy*, 192:186–193, 1998.

[272] D. Young and A. J. Gray. Cell identification in differential interference contrast microscope images using edge detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 1, pages 133–142. BMVA Press, 1996.

[273] W. Yurcik and C. Liu. A first step toward detecting SSH identity theft in HPC cluster environments: Discriminating masqueraders based on command behavior. In *Proceedings of the Fifth IEEE International Symposium on Cluster Computing and the Grid (CCGrid)*, volume 1, pages 111–120. IEEE, Washington, DC, USA, 2005.

[274] A. Zaia, R. Eleonori, P. Maponi, R. Rossi, and R. Murri. Medical imaging and osteoporosis: Fractal's lacunarity analysis of trabecular bone in MR images. In *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, pages 3–8. 2005.

[275] A. Zell. *Simulation Neuronaler Netze*. Addison-Wesley, 1994.

[276] F. Zernike. Beugungstheorie des Schneidenverfahrens und seiner verbesserten Form, der Phasenkontrastmethode. *Physica*, 1:689–704, 1934.

[277] H. Zhang, P. Morrow, S. McClean, and K. Saetzler. Incorporating feature based priors into the geodesic active contour model and its application in biomedical imagery. In *Proceedings of the International Machine Vision and Image Processing Conference (IMVIP)*, pages 67–74. IEEE, 2007.

[278] C. Zimmer, E. Labruyère, V. Meas-Yedid, N. Guillén, and J.-C. Olivo-Marin. Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: A tool for cell-based drug testing. *IEEE Transactions on Medical Imaging*, 21(10):1212–1221, 2002.

[279] F. G. Zöllner. *Enhancing Protein–Protein Docking by new approaches to Protein Flexibility and Scoring of Docking Hypotheses*. Ph.D. thesis, Faculty of Technology, Bielefeld University, Germany, 2004.

Notation and Symbols

$!$	factorial
(a, b)	tuple consisting of the variables a and b
$[a, b[$	interval: $a \leq x < b$
$[a, b]$	interval: $a \leq x \leq b$
$]a, b]$	interval: $a < x \leq b$
$\langle \underline{a}, \underline{b} \rangle$	dot product of the two vectors $\underline{a}$ and $\underline{b}$
∇	gradient operator
$\vee$	point-wise maximum
$\wedge$	point-wise minimum
$\{a, b, c\}$	set comprising the elements a, b and c
$\lvert \cdot \rvert_1$	city block norm
$\lvert \cdot \rvert_2$	Euclidean norm
$\lvert \underline{A} \rvert$	determinant of a matrix $\underline{A}$
$\underline{A}$	mixing matrix of the independent component analysis
A_j	area dependent segmentation error reflecting the fraction of incorrectly segmented pixels of a segment j
A_j^{diff}	number of pixels that a segment j and its reference segment differ in (Hamming distance)
A_j^{man}	area of a manually extracted cell mask that corresponds to segment j
A_{pq}	Zernike moment of order (p, q)
$\mathcal{B}(n, p)$	binomial distribution of n experiments with a success probability of p
$\underline{C}_{\Phi(\underline{x})}$	covariance matrix regarding observations of the vector $\underline{x}$, which were transformed into a space $\mathcal{H}$ using the map $\Phi(\underline{x})$
$\underline{C}_{\underline{x}}$	covariance matrix with respect to observations of the vector $\underline{x}$
D^F	fractal dimension
D^T	topological dimension
Δ_{max}	maximum distance from a cell marker that a snake can grow

E	energy
$\underline{E}$	transformation matrix of the principal component analysis
$E_{\text{ao}}\big(c(s)\big)$	snake energy that represents information from the cell membrane image (result of an algebraic opening)
$E_{\text{cont}}\big(c(s)\big)$	energy describing the continuity of a snake
$E_{\text{curv}}\big(c(s)\big)$	energy representing the curvature of a snake
$E_{\text{dist}}\big(c(s)\big)$	energy reflecting a snake's distance to the corresponding cell marker
$E_{\text{ext}}\big(c(s)\big)$	external energies of a snake
$E_{\text{int}}\big(c(s)\big)$	internal energies of a snake
$\bar{E}^{\text{seg}}$	mean of the similarity measure E_j^{seg}
E_j^{seg}	similarity measure reflecting the segmentation error
$E^*_{\text{snake}}\big(c(s)\big)$	energy functional describing a snake $c(s)$
H	measure for fractal properties
I	grey-scale image
$\underline{I}$	identity matrix
$\underline{K}$	Gram matrix (kernel matrix)
Λ^F	fractal lacunarity
$\mathcal{N}\big(\mu, \sigma^2\big)$	normal distribution with mean μ and standard deviation σ
$\Phi\big(\underline{x}\big)$	map from the original input space $\mathcal{X}$ to an alternative space $\mathcal{H}$
R^*	maximum category radius
$R_M^{\delta}(I)$	morphological reconstruction by dilation
$R_M^{\varepsilon}(I)$	morphological reconstruction by erosion
$R_i(t)$	radius of the category i
$V_{pq}(r, \alpha)$	Zernike polynomial of order (p, q)
a_j	length of the semimajor axis if the cell mask corresponding to segment j is approximated by an ellipse
α	weight of the snake energy $E_{\text{cont}}\big(c(s)\big)$; choice parameter of ART and ARTMAP nets; significance level
α_i	Lagrange multiplier
β	weight of the snake energy $E_{\text{curv}}\big(c(s)\big)$
c	speed of light

$c(s)$	a curve representing a contour, e.g. a snake
$c(t)$	classification results at the time step t
d	degree of the polynomial kernel
$\bar{d}$	mean of the segmentation error d_j
δ_0	base value of the energy weight $\delta\Big(E_{\text{dist}}\big(c(s)\big)\Big)$
$\delta\Big(E_{\text{dist}}\big(c(s)\big)\Big)$	weight of the snake energy $E_{\text{dist}}\big(c(s)\big)$
$\delta_M^{(1)}(I)$	elementary geodesic dilation
$\delta_M^{(n)}(I)$	geodesic dilation with n iterations
$\delta_S(I)$	morphological dilation of an image I using the structuring element S
d_j	shape dependent segmentation error which is based on the distance from a segment j to its reference contour
$d_j^{\max}$	maximal distance from a segment j to its reference contour
d_{m}	approximate diameter of the cell markers
e	entropy
$\underline{e}_j$	eigenvector
ε	eccentricity
$\varepsilon_M^{(1)}(I)$	elementary geodesic erosion
$\varepsilon_M^{(n)}(I)$	geodesic erosion with n iterations
$\varepsilon_S(I)$	morphological erosion of the image I using the structuring element S
η	learning rate
$f(\underline{g}_i)$	fitness of an individual i with the genome $\underline{g}_i$
f^{k}	fraction of images showing known protein locations that are classified as possibly unknown (before retraining)
$f_m(w_j)$	weight modification function
f^{u}	fraction of images showing unknown protein locations that are classified as possibly unknown (before retraining)
$g\Big(\Big\lvert\nabla I\big(c(s,t)\big)\Big\rvert_2\Big)$	edge detector function
γ	parameter of the RBF and the sigmoid kernel
γ_0	base value of the energy weight $\gamma\Big(E_{\text{dist}}\big(c(s)\big)\Big)$

$\gamma\Big(E_{\text{dist}}\big(c(s)\big)\Big)$	weight of the snake energy $E_{\text{ao}}\big(c(s)\big)$
$\gamma_S(I)$	morphological opening of the image I using the structuring element S
$\underline{g}_i$	genome of an individual i
h	Planck's constant
$\text{id}_i(j)$	mean intensity difference for pixels with distance j
$\iota(I, I^{\text{m}})$	minima imposition using the mask image I and the marker image I^{m}
$k\big(\underline{a}, \underline{b}\big)$	kernel of the two vectors $\underline{a}$ and $\underline{b}$
κ	Euclidean curvature; parameter modifying the influence of $E_{\text{dist}}\big(c(s)\big)$; weight modification exponent
l	length of the linear structuring elements used for the determination of cell membrane pixels
λ	wave length; length of line segments
λ_j	eigenvalue
m	median
μ	mean
$\underline{\mu}$	mean vector
$\underline{\mu}_i(t)$	centroid of a category i
μ_n	statistical moment of order n
μ_{pq}	central moment of order (p, q)
$\underline{n}$	normal vector
n_{bg}	maximum number of iterations for the geodesic erosion of the image background
n_{c}	number of localised cells
n_{m}	number of determined cell markers
ν	frequency; parameter of ν-SVCs; parameter of one-class SVCs
$\phi_S(I)$	morphological closing of the image I using the structuring element S
ρ	vigilance parameter
$\varrho_S(I)$	self-complementary top-hat applied to the image I using the structuring element S
r_i	correlation of the segmentation error with a specific feature x_i
s	smoothness

$\mathrm{sgn}(x)$	sign function
σ	standard deviation
τ	threshold regarding $\tilde{z}_{\mathrm{min}}^{F2}(t)$ for considering a presented input vector $\underline{x}(t)$ as being known
τ_2	threshold regarding $\tilde{z}_{\mathrm{min}}^{F2}(t)$ for detecting new protein locations
τ_2^{opt}	optimal value of τ_2
τ_{bg}	threshold for the geodesic erosion of the image background
τ_{m}	threshold for the recognition of cell membrane pixels
τ_w	threshold for rejecting features
θ	bias
$\mathrm{tr}(\underline{A})$	trace of a matrix $\underline{A}$
u	uniformity
$\underline{w}$	weight vector
$\underline{w}_i^{F2}(t)$	weight vector of an $F2$ neuron i at the time step t
x_i	single input (feature)
$\bar{x}_i$	mean of a feature x_i
$\underline{x}(t)$	input vector (feature vector) at the time step t
$y_i^{F2}(t)$	output of an $F2$ neuron i at the time step t
$z_i^{F2}(t)$	activation of an $F2$ neuron i at the time step t
$\tilde{z}_i^{F2}(t)$	alternative measure for the activation of the $F2$ neuron i, which corresponds to the distance from an input vector $\underline{x}(t)$ to category i
$\tilde{z}_{\mathrm{min}}^{F2}(t)$	minimum distance to all existing categories that indicates the degree of knowledge of an input vector $\underline{x}(t)$

Index