Functional proteomics:

Generation and analysis of cDNA-encoded proteins

Susanne Gräslund

Stockholm 2002

ABSTRACT

Recent advances in genomics have led to the accumulation of vast amounts of data about genes. However, it is the proteins and not the genes that sustain function, a fact which makes the proteins the keys to understand biology. Unlike the genome, which is a fairly constant entity, interesting biological and medical questions relate to the dynamic world of the proteomes expressed in different cell types and under different conditions. Proteomics, the large-scale analysis of proteins, now aims to identify and map entire proteomes and the challenge of studying proteins on a global scale is driving the development of new technologies for systematic analysis of protein function.

The first step towards gene characterization is to obtain its complete coding sequence. Here, a method to retrieve the upstream coding sequence of a gene with a partially known downstream sequence is described. The concept is based on a polymerase-chain reaction (PCR)-assisted biotin-capture method performed directly on poly(A)+ RNA to generate a full-coding sequence of a gene. The method was applied to the gene TSG118, of which only a partial sequence was known and for which a full-length clone was not found in the available cDNA libraries.

In functional analysis of proteins, information of spatiotemporal localization at cellular and subcellular level is important information, since it provides clues to function and suggests further experimentation. This thesis describes the development of systems, aimed at large-scale localization of cDNA-encoded proteins based on the generation of highly specific polyclonal antibodies. Significant efforts have been invested in the development of robust and general expression systems, suitable for high-throughput production of cDNA-encoded proteins in E. coli. A single-vector concept was first developed and evaluated for the production of 55 cDNA products from a mouse testis library. More than 90% of the expressed gene products were recovered with good yields. Subsequently, a dual vector concept was created in order to allow a more stringent procedure for affinity enrichment of the antibodies to be used for functional annotation. Antibodies generated by the described approach have been used to characterize genes encoding members of a protein complex and a tektin protein (APC/C and Tekt1).

An expression system employing a novel tailor-made affinity handle was developed and evaluated. A Z-affibody, showing binding capacity toward protein A, had earlier been selected from a library constructed by combinatorial mutagenesis of a protein A domain. It was now used as an affinity handle enabling efficient recovery on protein A-Sepharose, a robust and well-documented chromatography medium. The system was used for production of cDNA-encoded proteins and in addition, two convenient affinity blotting procedures were developed to allow screening of expression efficiencies. The robustness and convenience of the presented expression system should make it suitable for various high-throughput protein expression approaches.

An effort to create single vector three-frame expression systems, allowing expression of an inserted gene in three reading frames, is presented. The aim was to create a combined cloning and expression vector, in order to simplify cloning and protein expression procedures. Two vectors were constructed and although the systems would need further optimization to be used in high-throughput protein production, the principle of single vector three-frame expression was demonstrated.

Key words: functional genomics, functional proteomics, gene characterization, E. coli expression, affinity purification, immunolocalization, mouse testis, spermatogenesis, rabbit antibodies, affibody, tektin, Anaphase-promoting complex or cyclosome.

ISBN 91-7283-234-7
"Den mätta dagen, den är aldrig störst.
Den bästa dagen är en dag av törst."

Karin Boye
LIST OF PUBLICATIONS

This thesis is based on the papers listed below. They are referred to in the text by their Roman numbers.


# TABLE OF CONTENTS

## INTRODUCTION

1. Historical perspective 2

## Functional analysis of genes and gene products 3

2.1. Genomics – getting the genes 3

2.1.1. Gene accessibility 4

2.1.2. Genetic variability 6

2.2. Transcriptomics 6

2.2.1. *In situ* hybridization 7

2.2.2. DNA Microarrays - Expression profiling 7

2.2.3. Blotting techniques 8

2.3. Proteomics 9

2.3.1. Protein identification 10

2.3.2. Protein-protein interactions 12

2.3.3. Protein localization 13

2.3.4. Protein structure 16

2.3.5. Global proteome studies 17

## Recombinant protein expression and purification 17

3.1. Hosts for expression 18

3.2. Affinity tags 20

## Antibodies as tools in proteomics 22

4.1. Recombinant antibodies 23

4.2. Antibody arrays 24

## Present Investigation 26

5. Recovery of upstream cDNA sequences (I) 26

6. Development of high-throughput expression systems (II-V) 28

6.1. A high-throughput protein expression system 29

6.2. A high-stringency dual concept 33

6.3. A novel affinity tag allowing protein A-based recovery 36
6.4. Single vector three-frame expression systems 38

7. Functional analysis based on protein expression (VI-VII) 42
   7.1. Characterization of the human APC1 42
   7.2. Tekt1 – the first mammalian tektin 43

8. Concluding remarks and future perspectives 45

ABBREVIATIONS 48

ACKNOWLEDGEMENTS 49

REFERENCES 50

ORIGINAL PAPERS (I–VII) 60
INTRODUCTION

It is often postulated that we are in the middle of a biological revolution, or even a paradigm shift, although we cannot now see it. The full extent of this will only be visible to us when time has given a little perspective on the present. Nevertheless, it is true that we live in a very exciting time for bioscience and I believe that in the future, people will look back on the time around the turn of the millennium as the time when the secrets of the human genome started being unveiled.

With the creation of the Human Proteome Organisation (HUPO) (Agres, 2001), and the Human Proteome Project we have entered the proteomics era. Mapping single proteins and signaling pathways has been done for a long time, but now the aim is to map the entire proteomes of a variety of species, including *H. sapiens*. The term proteome was first coined in 1994 by an Australian postdoctor named Marc Wilkins, and refers to the total set of proteins expressed in a given cell at a given time. Compared to the Human Genome Organisation (HUGO), that together with several private efforts has brought forth the entire genome sequences of a number of species by DNA sequencing, studying and characterizing proteins cannot be done equally straightforward. Since proteins are more complex than DNA, there does not exist one single method to be used on a broad scale for functional characterization of proteins, but rather several methods and approaches that complement each other to help us reach the goal. The challenge of studying proteins in a global way is also driving the rapid development of new technologies needed for this immense task to be completed (Lee, 2001).

DNA can be described as the memory of a cell, and is used to store information, almost without any modifications at all, for the future. It is built up by only four different elements, i.e. nucleotides, that are very similar in their chemical properties. Proteins however, are the active agents of a cell, the ones who sustain function and gives every cell type its unique character. Proteins are made of 20 amino acids that exhibit a variety of different biochemical properties. In addition, proteins fold and assemble subunits to give complex three-dimensional structures, which also can change, while the protein exerts its function. Furthermore, it is also apparent that the paradigm of one gene encoding a single protein is no longer tenable. Processes such as alternative splicing, RNA editing, post-translational modifications increase the functional complexity of an organism far from what is indicated in its genome sequence alone. Unravelling this complexity will be a major challenge of the post-genomic era.
1. Historical perspective

The understanding of the mystery of life has been a central driving force for many scientists and philosophers during the history of humankind. Not until almost the last century did we get a hint of the truth by the discovery of genes (Gregor Mendel, 1865) and the common molecular elements of earthly life - DNA, RNA and proteins. Studies by Avery et al in 1944 suggested that DNA was the genetic material and only a few years earlier, in 1941, Beadle and Tatum had established a connection between genetic material and phenotypic traits (protein activity). Watson and Crick discovered the double helical structure of DNA in 1953 and thus gave us the answer to how the genetic material is organized. In the 60s the process whereby genes are translated into proteins via messenger RNA was deciphered by Holley, Khorana and Nirenberg - a process which is now called the central dogma of bioscience.

![Diagram of the central dogma]

During the last decades, there has been an explosion of new discoveries and new technologies to make more discoveries. DNA sequencing (Maxam and Gilbert, 1977; Sanger et al., 1977), and the polymerase chain reaction (PCR) method to amplify genetic material (Saiki et al., 1985; Mullis and Faloona, 1987), are only examples of methods that have contributed to the biological revolution. The large number of Nobel prizes awarded to the field of bioscience also bears witness of the importance and influence these discoveries have had on our understanding of the biological mechanisms of life.
2. **Functional analysis of genes and gene products**

The methodology of molecular biotechnology is very much about making visible what can not be seen with the keen eye, for example analysing the sequence of a gene or the structure of a protein. The following sections contain an overview of a number of methods used to functionally characterize genes, transcripts and gene products, developed in the last decades in the field of bioscience.

2.1. **Genomics – getting the genes**

In the past five years, more than 30 bacterial and six eukaryotic genomes (Table 1) have been completely sequenced and reported, the first being that of the bacterium *Haemophilus influenzae* (Fleischmann et al., 1995). In addition, the two eukaryotic sequences of *Homo sapiens* (Venter et al., 2001) and *Mus musculus* are now available to a large extent (Legrain et al., 2001). The availability of the complete genomes of course gives us a much better overview of the organisation of life and also an extensive starting material for the proteome project; to decipher the functions of all genes and gene products.

**Table 1.** Complete eukaryotic genomes that have been reported to GenBank by December 2001.

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome size (Mb)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>125</td>
<td>(Arabidopsis Genome Initiative, 2000)</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>100</td>
<td>(C. elegans Sequencing Consortium, 1998)</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>120</td>
<td>(Adams et al., 2000)</td>
</tr>
<tr>
<td><em>Encephalitozoon cuniculi</em></td>
<td>2.9</td>
<td>(Katinka et al., 2001)</td>
</tr>
<tr>
<td><em>Guillardia theta</em></td>
<td>0.5</td>
<td>(Douglas et al., 2001)</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>12</td>
<td>(Goffeau et al., 1996)</td>
</tr>
</tbody>
</table>
Determining gene function through genomics typically starts from a query of a database. Access and interpretation of the large amounts of available sequence material often require sophisticated bioinformatics software tools to select the desired sequences. Once the chosen genes are found in the virtual world the next requirement is to get the actual genetic clones of the genes.

2.1.1. Gene accessibility

The genetic material, DNA, occurs naturally in chromosomes and are ordered in coding regions, exons, which are interrupted by large non-coding regions, introns. Prokaryotes, which lack introns, have a higher percentage of coding regions than eukaryotes. In the human genome less than 5% of the sequence is believed to be actual genes and these are often interrupted by several introns of varying sizes. The desired genetic material, being genes or gene fragments encoding the proteins to be investigated, could potentially be obtained by PCR amplification of chromosomal DNA. This method is suitable only if the goal is to amplify single exons or if PCR splicing could be applied to join a few exons and exclude the introns (Horton et al., 1989). An obvious advantage to use chromosomal DNA as source to gain the desired genetic material is that chromosomal DNA is readily available and that all genes are surely present.

Since less than five percent of the genomes of higher eukaryotes actually codes for genes, a lot of the sequencing efforts for those species have been focused on so called expressed sequence tags (ESTs). All expressed genes are transcribed to a messenger RNA (mRNA) before translation to a protein. Recovery and reverse transcription of the mRNA results in complementary DNA (cDNA) that represents the entire code of a gene, without introns which are spliced out when the mRNA is formed. An EST is such a cDNA fragment and could either correspond to part of the coding sequence or alternatively to untranslated 3’ or 5’ mRNA sequences. The amount of EST sequences reported to the public databases is very large and the problem is now to sort and organize the information into a larger picture showing the outline of the active genome. The total mRNA from a cell converted to cDNA represents all expressed genes at that time. This is called a cDNA library and consists of both full-length clones and smaller tags, usually subcloned into a library vector. Many genes are believed to be expressed in several splice variants which theoretically would all be included in a cDNA library. Desired cDNA clones can thus be amplified from such a library, and this method to get physical clones has been widely used lately. Since all members in the library are subcloned into the
same library vector, general primers can be used to amplify all clones. The drawback of cDNA libraries is the lack of guarantee that your clone of interest actually is present in the library. Also, sequence errors in the clones are not uncommon. Today a large number of cDNA clones can be bought as IMAGE (Integrated Molecular Analysis of Genomes and Their Expression) clones from the IMAGE consortium (Lennon et al., 1996). This consortium was founded to array, sequence and map a collection of cDNAs, representing among others all human genes, and to distribute these to the public. Several public and private full-length cDNA clone collections, with an increasing number of clones, are presently being created since these are seen as the sources of genes for various proteomics efforts.

Another method for the generation of cDNA libraries is the reverse-transcriptase – PCR (RT-PCR) method which can be used for isolation of a specific gene or gene fragment directly from total RNA or mRNA pools. Instead of making a cDNA library of the total RNA of a cell and then screen the library, only the desired clones are caught and reversely transcribed to cDNAs. Total RNA from several tissues and cell types are pooled together and used as starting material. Clone specific primers are thus designed and used for the reverse transcription and amplification. This has proved to be a very efficient method to gain cDNA clones, but with the drawback that specific primers have to be synthesized for every single clone (Agaton, 2002).

Another method to obtain genetic material would be to synthesize oligonucleotides and assemble them in vitro. When wanting to make specific base alterations in genes, this is an efficient method. The first oligonucleotide is preferably biotinylated and allowed to bind to a streptavidin-coated paramagnetic bead. Subsequently the following oligos are added, one by one, allowed to anneal and then ligated to the growing construct (Ståhl et al., 1993; Paper V). Gene fragments consisting of several hundred nucleotides can be assembled in this manner, but with the obvious drawback of cost for oligonucleotide synthesis.

Almost all these methods require PCR amplification and reverse transcription to gain the genetic material. However, since reverse transcriptase and most polymerases lack proofreading activity, care must be taken to limit the number of amplification steps in order to minimize the probability of introducing unwanted mutations.
2.1.2. Genetic variability

When the complete genomes of several species have been fully reported, the sequencing efforts have somewhat shifted focus. Since so large parts of the genomes are quite homologous between species and more evidently between two individuals of the same species, present interest is focused to the small differences on the genetic level. Although these differences seem small, the cumulative effects of them are responsible for the phenotypic differences between individuals. These variations often consist of only a single base that has been mutated, and there are some common hot spots on the genome that are more likely to differ between people. These sites are called single nucleotide polymorphisms, SNPs, and large efforts are now being put into the mapping of these sites and analyzing the impact they have on the phenotype.

SNPs are frequently present in the human genome with a density of at least one common SNP per kilobase pair (Lai et al., 1998) making a total of about 3 million. The SNP consortium (TSC), founded 1999 by 13 pharmaceutical companies and the Wellcome Trust, has set out a goal to discover a minimum of 300,000 SNPs and ensure public accessibility of the results (Lai, 2001). The high-density genetic map of the SNPs, showing the linkage between phenotypic and genotypic data, is likely to become very important in diagnostics and medical treatment in the future.

2.2. Transcriptomics

The term transcriptome is usually defined as “the total transcript complement of a genome” and transcriptomics is the discipline for transcriptome studies. It is important to study expression and expression patterns to gain information about the differences between cell types and the impact of different environments on cell behavior. Differences in expression patterns can also suggest functions of proteins that are differentially expressed under various conditions. We know now that there is no strict quantitative correlation between the mRNA levels and the amount of translated protein (Gygi et al., 1999 and 2000), but still, changes in transcription give us information on how the cell responds to inner and outer stimuli. Also, the presence of a transcript in a cell indicates that the corresponding gene is active, at least to a certain level, in that cell.
2.2.1. In situ hybridization

In situ hybridization is a valuable and well-established method for localization of gene expression on tissue level (Dagerlind et al., 1992). The hybridization method exploits the specific recognition of complementary DNA and RNA strands. A DNA or RNA probe complementary (anti-sense) to the transcript (mRNA) of a gene of interest is generated and labeled either radioactively or fluorescently. Tissue sections are prepared and incubated with the probe, which specifically recognizes cells harboring the target transcript.

Cellular maps of gene expression within cell types and tissues can be provided by in situ hybridization (Bankfalvi and Schmid, 1994). Advantages as compared with immunohistochemical techniques (see section 2.3.3) are the sensitivity that allows for detection of down to tens of mRNA molecules (Harris et al., 1996) and the ability to detect activity of a gene independently of the final protein product. Furthermore, it is easy to envision large-scale generation of probes covering complete transcriptomes for in situ hybridization techniques.

A major inherent limitation is due to the target being the transcript, mRNA, whose location does not provide subcellular localization for the gene product or even a proof of the existence of the protein. Another problem is the occurrence of non-specific hybridization and thus falsely positive results. To circumvent this, carefully performed control experiments are required.

2.2.2. DNA Microarrays - Expression profiling

The development of high-density array technologies provides the opportunity to comprehensively and efficiently survey the gene expression pattern of different cell types under varying conditions (Harrington et al., 2000; Ferea and Brown, 1999). DNA hybridization array technologies, for the simultaneous monitoring of whole genome activity in a single assay, appeared on the genomic arena less than a decade ago. Originally derived from the work by Ed Southern on detection of specific sequences among DNA fragments (Southern, 1975) the technique utilizes the interaction between complementary DNA or RNA strands, hybridization, for the detection of genes. Each array consists of a reproducible pattern of thousands of different probe DNAs, primarily PCR products or oligonucleotides, attached to a solid support which is usually glass (Schena et al., 1995). Fluorescently labeled RNA or DNA prepared from mRNA is hybridized to complementary DNA on the array and then detected by laser scanning.
Several methods have been described for producing microarrays, and two basic types are most commonly used; spotted arrays where pre-synthesized DNAs are printed onto glass slides and high-density oligonucleotide arrays on which sets of oligomers are synthesized in situ on glass wafers (Harrington et al., 2000). Spotted arrays can be produced in-house whereas high-density arrays require advanced instrumentation and are commercially available. Microarray methods have been used in a variety of experiments including the analysis of gene expression in human cancer (DeRisi et al., 1996) and identification of yeast cell cycle-regulated genes (Spellman et al., 1998). One important information derived from expression profile analysis is the elucidation of different temporal patterns of gene regulation and these two examples show the use of this technique for the identification of gene activity that follows a particular pattern. By looking at the patterns of characterized genes, knowledge of their function can be used to indicate functions of uncharacterized genes with the same pattern of regulation (Iyer et al., 1999).

DNA hybridization technologies are used in many other applications as well, for example in mutational screens (Saiki et al., 1989) including large-scale SNP detection (Wang et al., 1998).

A limitation with expression profile analysis by hybridization microarrays, is the prerequisite to know the genes, whose regulation one is attempting to study. This limitation is of course diminishing as the availability of sequence information is increasing. Other limitations are the problems with unspecific hybridization, cross-hybridization and to achieve large enough samples for detection of even the rare transcripts (Duggan et al., 1999). Another limitation, which is inherent in the expression profiling method, is the measurement of mRNA levels as indicator of gene activity. This limitation is attributed to the inability to measure the real gene activity since it is first made real on the protein level and the amount of transcript in a cell can not be used to establish protein amounts (Haynes et al., 1998).

Even with limitations with respect to the real activity of the functional protein, there are still tremendous insights into the inner workings of a cell to be gained with these mRNA-based techniques. With the completion of the sequencing of several genomes, the analysis of global gene expression patterns is likely to play an important role in understanding biology.

2.2.3. Blotting techniques

In 1975 a blotting technique for detection of specific DNA molecules in a mixture of different DNA molecules was developed by Edward Southern (Southern, 1975).
The method was named Southern blotting after its founder and when it was followed by the development of a related technique for RNA detection the latter was somewhat humorously denoted Northern blotting (Thomas, 1980). A third method for the detection of proteins has been named Western blotting (Burnette, 1981) and will be further described in section 2.3.3.

The sample for Northern blotting is typically total cellular RNA from purified cells or tissue samples. The RNA is denatured and treated to prevent formation of secondary structures by base pairing, thereby promoting the existence of unfolded linear RNA. The components of the RNA mixture are separated according to size by gel electrophoresis followed by blotting (transferring) of the RNAs to a nitrocellulose or nylon membrane. The membrane is then exposed to a labeled DNA probe specifically recognizing the RNA molecule that is being studied. By autoradiography the bands recognized by the probe, thus corresponding to the RNA of interest, are stained on the membrane. Northern blotting provides a convenient way to study within which cell types or tissues a certain gene is expressed. Since the staining intensity is correlated to the amount of RNA, at least semi-quantitative information of RNA can be obtained with this method. But as with all methods monitoring transcript levels, it suffers from the inherent limitation that mRNA levels cannot be used to establish protein levels (Haynes et al., 1998).

2.3. Proteomics

Classical proteomics methods have often been applied to proteins in a case-by-case manner, slowly making progress in characterization of the proteomes. Now, with the large amount of available genomic data, high-throughput efforts studying proteomes in a global scale can be introduced. The entire field of proteomics covers many different aspects. First, all the different cell types have to be mapped and all proteins identified. Second, the temporal and spatial localization of every gene product as well as the relative abundance of every protein has to be determined. Third, the structure of every protein has to be solved and finally, the protein interactions for the entire cell have to be mapped giving us a picture of the complex network that makes up the cell. Although it will probably take mankind several decades to create a complete picture of the proteomes, already today significant insights into protein function, phenotypic patterns and the understanding of complex diseases can be gained by proteomics efforts.
2.3.1. Protein identification

If the members of a complex proteomic mixture are to be identified, they first have to be separated. The by far most commonly used method to achieve this separation is two-dimensional electrophoresis (2D-E) which resolves complex mixtures of proteins first by isoelectric point and then by size. The method has been widely used for more than 25 years and the coupling of it with mass spectrometry (MS) to identify the separated proteins has made it one of the most important methods in proteome analysis. However, there are a few limitations with 2D-E: sensitivity, resolution, reproducibility and the inability to monitor proteins that are very hydrophobic or very acidic or basic.

The sensitivity of 2D-E depends on the labeling techniques used for detection. Labeling could either be incorporated into the proteins in vivo or by staining the proteins after separation. One way to obtain in vivo labeling is to grow the cells in the presence of [35S]methionine (Freyria et al., 1995). Staining methods after separation employ reagents such as Coomassie Brilliant Blue, silver stain or fluorescent dyes. The labeling intensity is proportional to the amount of protein present in the mixture and so, very rare proteins are impossible to detect among the more abundant. This is a major limitation since the rare ones are likely to be an important group of proteins because they encompass many receptors, signal transduction and regulatory proteins.

The resolving power of 2D-E depends on the conditions in the gel and several methods have been developed to increase the number of proteins that can be separated. The first is to run multiple gels covering overlapping pH and/or molecular weight regions. A second way is to prefractionate the sample to reduce its complexity, thus adding a third dimension of separation. Using these methods, it is possible to resolve more than 10,000 proteins from a higher eukaryotic cell lysate (Wildgruber et al., 2000). However, when considering that the human proteome consists of several ten thousands of proteins, this resolution is far from being fully sufficient.

2D-E is a technology that involves a great deal of expertise and hands-on time to execute reproducibly. A standard set of running conditions has never been agreed upon (Fey and Larsen, 2001) which makes it difficult to compare individual 2D-E patterns from different laboratories in detail before the protein spots have been identified. The extensive development in identification methods by MS has been
important to counteract this problem. Also prefractonation, if used, might introduce additional variability into the results.

Another important limitation of 2D-E is the difficulty to separate proteins with extreme properties regarding hydrophobicity (e.g. membrane proteins) and charge. It has been recently estimated that approximately 30% of proteins are membrane proteins (Paulsen et al., 1998). Considering their key roles in signal transduction, cell adhesion, metabolite and ion transport and the fact that they are often the target for drug interactions, it remains critical to solve this problem (Santoni et al., 2000). The use of specific detergents has been reported (Wissing et al., 2000), and seems to be a promising approach in this aspect.

After the 2D-E separation the individual spots are isolated and the proteins identified. The latter was traditionally accomplished by N-terminal sequencing (Edman, 1950) internal peptide sequencing (Rosenfeld et al., 1992), immunoblotting or co-migration (Honore et al., 1993). Currently the most efficient techniques for identification are based on mass spectrometry (MS) with matrix-assisted laser desorption-ionisation time-of-flight (MALDI-TOF) as the preferred technique. When using MS the protein spots are fragmented chemically or enzymatically to yield a peptide digest. These peptides are then analyzed by MS to obtain a mass spectrum of peptides. By comparing these masses with those obtained from databases, the identity of a certain protein can be established (Patterson and Aebersold, 1995). Although this peptide mapping approach provides a fast and simple way to identify proteins it suffers from several limitations, the most evident being the need for the protein to be already present in the databases to be identified. Moreover, the 2D-E resolution may not have achieved a complete separation, which gives mass spectra containing peptides from more than one protein. Furthermore extensive post-translational modifications can result in peptide masses not correlating with those obtained from the databases. Also the existence of errors in the databases poses a problem in protein identification analysis.

The development in this area is very extensive and several systems aimed for high-throughput protein analysis have been put together. Tandem mass spectrometry apply a second analysis step following the peptide mapping where selected peptide bands are dissociated to produce mass spectra characteristic for the peptide sequence (Figeys et al., 1996). Mass spectrometry can also be used to study post-translational modifications in a high-throughput manner (Wilkins et al., 1999).
2.3.2. Protein-protein interactions

To produce a specific phenotype of a cell, genes are expressed in a certain pattern and the resulting proteins interact with each other in the network comprising the living cell. These interactions also include those between proteins and nucleic acids. Protein-protein interactions can be in the form of modification processes that modify the functions of one or more of the involved proteins. Signaling pathways and complex formations are good examples of this mechanism. Complexes may be more or less stable; depending of the function that can for example be structural or actively carrying out a function. Several selection strategies are being used to find proteins and other molecules that bind to each other in a specific way, many taking the form of a bait and prey system where binding pairs are identified in a single selective step.

Interactions between proteins and nucleic acids have classically been studied using biochemical techniques, co-precipitation and co-fractionation by chromatography. Another method that has been extensively used during later years is the two-hybrid systems (2HS) which was originally developed for yeast (Fields and Song, 1989; Chien et al., 1991). Two-hybrid systems are based on the separation of a protein into two non-functional domains, which are then fused to binding candidates. If the fusion partners bind to each other, the function of the separated protein will be restored. This has to be reported in some way so that the cells harboring binding pairs can be identified and studied. 2HS has traditionally been performed in yeast cells using the GAL4 protein which is a transcriptional activator, to activate a certain reporter gene. An extension of the original system that allows the analysis of three interacting molecules was realized in the three-hybrid systems (3HS) (Zhang and Lautar, 1996).

The two-hybrid system can also be used on a genome-wide scale to generate protein interaction maps for complete proteomes. However, some inherent limitations exist in the original system. First, the interaction between proteins is monitored in the nucleus, which imposes limitations for studying proteins that cannot enter the nucleus. Second, the studied proteins can themselves be transcriptional activators or repressors. Finally, yeast cells also have limitations for the analysis of mammalian proteins, regarding extensive post-translational modifications.

Another system that has been launched in later years is the dihydrofolate reductase (DHFR)-based selection in bacterial cells (Pelletier et al., 1999). Here, an enzyme
is separated into two parts making it non-functional. By fusing the two parts to potential binding proteins, the function of DHFR can be restored if the parts are reunited. Growing the cells in an environment where the capacity to produce DHFR is crucial for survival, cells containing binding pairs can be selected and identified. This system can be used both for target versus library and library versus library selections.

Phage-display systems have also been used for studies of protein interactions. The most common methods being selectively infective phage (SIP) (Krebber et al., 1997) and selection and amplification of phage (SAP) (Duenas and Borrebaeck, 1994). The SIP and SAP concepts have both been successfully applied in model selections, but as far as we know, neither has been used for library versus library selections. Both rely on binding between two candidate proteins to bring together separated domains of the phage pIII protein to restore its ability to infect bacteria. Both methods are hampered however, by there being a small window where the ratios of bait and prey are optimal. In practice, this means there is likely to be a strong bias for the selection of binding pairs whose expression best fulfil the criteria raised by the system (Holt et al., 2000).

Interactions between proteins and nucleic acids can alternatively be studied using an in vitro evolution process called systematic evolution of ligands by exponential enrichment (SELEX) (Ellington and Szostak, 1990; Tuerk and Gold, 1990). The method is based on the isolation of DNA sequences that interact with proteins from large libraries of random sequences. Successive rounds of binding, partitioning and amplification are performed resulting in the isolation of oligonucleotides with specific affinity to the investigated protein, so called aptamers. To date, aptamers that bind to small organic molecules, carbohydrates, amino acids, peptides and proteins have been identified (Green et al., 2001). A variant of the classical SELEX is the genomic SELEX (Singer et al., 1997) in which the sequence library is derived from the genome of an organism. This approach could be envisioned as a method for rapid genome-wide screening of interactions between proteins and nucleic acids. The most evident drawback of the SELEX methods is that they study the interactions in vitro where the conditions may differ from those in vivo (Shultzaberger and Schneider, 1999).

2.3.3. Protein localization

The eukaryotic cell has extensive internal structure and compartmentalization in which organelles and other structural components perform different specialized
functions. Depending on the spatial localization of a cell within an organism, a specific cellular phenotype is adopted in which a certain subset of the genome is actively transcribed. Also, the proteins expressed from the genome in a cell vary considerably with its spatiotemporal behavior. Localization information provides clues to function, based on knowledge about known structures and developmental processes, and is thus important to the annotation of gene function. Localization also provides a basic knowledge from which additional experiments can be designed (Fields, 1997). There are several methods for analyzing protein localization, including tagging with reporter genes and immunolocalization techniques.

Tagging a gene with a reporter is an efficient way to determine the temporal expression and localization of specific proteins within a cell. Two reporters have been extensively described, based on β-galactosidase (Burns et al., 1994; Wach et al., 1994), and the green fluorescent protein (GFP) (Nabeshima et al., 1997). The tagged genes can then be detected by fluorescence or enzymatic substrate conversion.

A gene can also be expressed in situ by an expression system in which the target gene is produced as a fusion protein with the reporter tag. This has been implemented with GFP in budding yeast for determination of the subcellular localization in living cells by fluorescence microscopy (Niedenthal et al., 1996). When compared to immunolocalization described below in this section, this method has the advantage that the protein is detected directly without the need for secondary reagents, such as antibodies, that can be sterically hindered. The possibility to study living cells is also enabled in this method, whereas for immunolocalization the cells have to be fixated. On the other hand, fusion to GFP may influence localization and might disturb the target protein from correctly exerting its function or even making it non-functional (Burns et al., 1994). Furthermore, expression of a recombinant fusion protein at a high level and the fact that the fusion protein has to compete with its native counterpart may also cause disturbances and even cell death.

Another approach is to tag the protein with a peptide tag for which an antiserum has been raised (Surdej and Jacobs-Lorena, 1994; Jarvik and Telmer, 1998). Such peptide tagging has the advantage of the requirement of only one antiserum to detect any protein, but the drawback is that proteins can only be detected one by one. As compared with the two previously mentioned protein tags, the peptide-epitope tag is not directly detectable, but needs a secondary reagent. However,
genome-wide tagging can be employed on smaller organisms such as yeast (Burns et al., 1994) and fruit fly (Spradling et al., 1995), but has so far not been compatible with mammals.

A powerful method for detection of specific proteins in a complex mixture is Western blotting or immunoblotting (Burnette, 1981). The method is analogous to Southern and Northern blotting described in section 2.2.3, with the difference that proteins are detected instead of nucleic acids. The first step of Western blotting is separating the protein sample by size using SDS-polyacrylamide gel electrophoresis (SDS-PAGE). Proteins are then transferred to a membrane, either by heat or electroblotting. Subsequently, the specific protein is then detected on the membrane by an affinity reagent (e.g. antibodies) specifically recognizing the target protein. A secondary antibody, which is enzymatically, fluorescently or radioactively labeled, and binds to the first one is then added. Staining is achieved by a method corresponding to the label type. Western blotting can be used in a number of applications including gene expression analysis and cell type and tissue screening for localization information, depending on which cell and tissue types that are available. Subcellular localization could potentially also be performed by blotting protein extracts from organelles or other subcellular structures. Most importantly, Western blotting is a very valuable initial expression screening method (Jörgensen et al., 1998) preceding more labor-intensive localization techniques such as immunolocalization.

To obtain information about the subcellular localization of proteins immunohistochemical methods have proven efficient. The base for these methods is the generation of an antiserum highly specific towards the target protein. Immunolocalization can be performed on either purified cells or tissue sections fixed on glass slides and incubated with the specific antiserum. Incubation with a labeled second antiserum can reveal a staining pattern that can be detected with various microscopy methods, for example light, immunofluorescence (IF) and confocal microscopy.

A potential problem with immunolocalization using either light or IF microscopy is that the pattern will be a superposition of all the layers in the cell sample. To get the picture of a single plane of focus, confocal microscopy can be used instead, which results in sharper images (White et al., 1987). To generate three-dimensional images, serial sections of fluorescent images can be put together. For suborganelle localization, the resolution of confocal microscopy limited by the wavelength of
light, may not be enough. Coating the sample with gold particles followed by immunoelectron microscopy might solve the problem (Geuze et al., 1981).

An obvious limitation with immunolocalization is the dependence of the accessibility of the native protein in the cell. Another detection problem is that the fixation procedure might destroy the structure in which the studied protein is involved or even its antigenicity. Today, it is maybe not envisionable to generate antisera towards all proteins in order to perform proteome-wide analyses, as it would indeed be a costly project. Techniques to circumvent the problem with immunizations, by using combinatorial protein chemistry and in vitro selection methods, are being developed (Persic et al., 1999; Krebs et al., 2001).

2.3.4. Protein structure

The knowledge of the structure of a protein is a first step to understand its function. The rapidly increasing number of three-dimensional structures solved, alone or in complex with other molecules, have contributed enormously to the understanding of how the order of amino acids is linked to the structure and ultimately to the function in catalysis, binding etc. Over the last decade, a large amount of structure knowledge has been provided by x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, of the exact three-dimensional structures of proteins, nucleic acids and their complexes and assemblies. Both methods require quite large amounts of pure full-length protein. Membrane proteins represent the most persistent bottleneck for all analytical methods, because they are water-soluble only in the presence of detergents and difficult to overproduce in quantities that are required for biophysical studies. Production of full-length mammalian proteins with a correct structure is a challenge, and requires the development of efficient expression systems.

Structural proteomics have set a goal to provide a comprehensive structural description of the protein universe. One approach focuses on structure analysis methodology and close co-operation with the genome sequencing projects. Efforts towards parallelization and automation of structure analysis by classical methods are unifying features of this approach, among others represented by the Berlin “Protein Structure Factory” (PSF) initiative in cooperation with the German Human Genome Project (DHGP) (Heinemann et al., 2000).

Another approach includes the possibility to predict the three-dimensional structure of a protein based on its amino acid sequence. However, the ability to predict the
structure of a protein is an intriguing problem in protein science. Even if the information about the final structure that a protein will adopt is embedded in the amino acid sequence, so far no one has been able to securely predict it. Structural proteomics now aims to determine a set of protein structures that will represent all domain folds present in the biosphere. At the level of building blocks of proteins, globular domains, it is believed that the number of folds is limited to no more than a few thousand (Heinemann et al., 2000). A structural library of all these could then be used as the basis for homology modeling of all remaining proteins. Computer-based methods for fold recognition are currently being developed in a number of laboratories (Bork and Eisenberg, 1998; Sowdhamini et al., 1996).

2.3.5. Global proteome studies

So far the described methods have been used primarily to study proteins in a case-by-case manner. But the proteins are not single entities “living separate lives” inside or outside cells. All the proteins are parts of a complex network of thousands of proteins and other molecules binding to or influencing each other in other ways. To get a picture of the whole network, global proteome studies have to be performed. Almost all methods described in the proteomics section can be applied on a global scale, and several attempts have been reported. The yeast two-hybrid system could possibly generate species-specific protein linkage maps by using cDNA libraries covering entire genomes, including T7 phage (Bartel et al., 1996), Hepatitis C (Flajolet et al., 2000), yeast (Ito et al., 2000), (Uetz et al., 2000) and C. elegans (Walhout et al., 2000). Also two-dimensional gel electrophoresis followed by MS analysis is a powerful tool in proteome-wide characterization and quantification. Furthermore, several new technologies are being developed for the examination of proteins using genomic formats. In the backwater of DNA microarrays, the idea has come forth to make protein arrays, with affinity reagents such as antibodies or other molecules, to detect single members of a complex protein mixture. The use of antibody arrays will be more thoroughly discussed in section 4.2 below.

3. Recombinant protein expression and purification

The efforts to characterize proteins from entire proteomes that are presently discussed (Anderson et al., 2001), clearly demonstrate the importance of high-throughput protein expression systems (Lesley, 2001; Albala et al., 2000). Ideally, such systems should not be optimized for a single protein, but allow for the
production and purification of a large number of proteins irrespective of their unique properties. As all life on Earth is based on the same common features, genetic material from different sources can be combined and introduced in several hosts for expression, a method called recombinant DNA technique. A wide variety of organisms are available as hosts depending on the requirements on the process and the quality of the protein product. These requirements may be to obtain full-length protein with a correct fold and required post-translational modifications.

The development of techniques for the cloning of genes and their subsequent expression in bacteria or mammalian cells in the early 1970s (Jackson et al., 1972; Cohen et al., 1973), opened up the possibility to produce large quantities of proteins that normally were available only in small amounts from natural sources or not at all. Human recombinant products have been produced in *E. coli* since the late 70s (Itakura et al., 1977) and since then a large number of medically important proteins have been produced. One of the most important examples is insulin, which is used to treat diabetes. The possibility to produce recombinant proteins for therapeutic use was a breakthrough since they became more available, cheaper and of better quality. In addition, the risk of virus contamination in connection to extraction of proteins from natural sources was avoided. Today recombinant methods are used for production not only of medically interesting proteins, but for proteins of use in many other fields as well.

### 3.1. Hosts for expression

Several types of hosts, both prokaryotic and eukaryotic, are used for recombinant protein production today (Table 2). The choice of which system to use depends upon many factors including protein size, structure, need for biological activity or post-translational modifications, possibilities of genetic engineering and downstream processing, yield and economy.

The dominating bacterial host for recombinant protein production is the gram-negative *Escherichia coli*, which is extremely well-documented and easy to work with. A vast number of expression strains and vectors are available today, as well as several promoter systems to regulate expression. Many proteins over-expressed in *E. coli* accumulate in the form of insoluble inclusion bodies (Hartley and Kane, 1988). However for many applications these inclusion bodies can be collected and the produced fusion proteins recovered later by a denaturation/renaturation protocol (Rudolph, 1990 and 1994; Fischer, 1993).
Table 2. Commonly used hosts for recombinant protein production.

<table>
<thead>
<tr>
<th>Host type</th>
<th>Example of organism</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gram-negative bacteria</td>
<td><em>Escherichia coli</em></td>
<td>(Baneyx, 1999; Hannig and Makrides, 1998)</td>
</tr>
<tr>
<td>Gram-positive bacteria</td>
<td><em>Bacillus subtilis</em></td>
<td>(de Vos et al., 1997)</td>
</tr>
<tr>
<td>Yeast</td>
<td><em>Staphylococcus carnosus</em></td>
<td>(Gellissen et al., 1992)</td>
</tr>
<tr>
<td>Yeast</td>
<td><em>Pichia pastoris</em></td>
<td>(Cregg et al., 2000)</td>
</tr>
<tr>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>(Gellissen et al., 1992)</td>
</tr>
<tr>
<td>Filamentous fungi</td>
<td><em>Aspergillus nidulans</em></td>
<td>(Devchand and Gwynne, 1991)</td>
</tr>
<tr>
<td>Insect cells</td>
<td><em>Drosophila melanogaster</em> cells</td>
<td>(McCarroll and King, 1997)</td>
</tr>
<tr>
<td>Plant cells</td>
<td>Tobacco cells</td>
<td>(Doran, 2000)</td>
</tr>
<tr>
<td>Mammalian cells</td>
<td>Chinese hamster ovary (CHO) cells</td>
<td>(Condreay et al., 1999; Wurm and Bernard, 1999)</td>
</tr>
<tr>
<td>Mammalian cells</td>
<td>African green monkey kidney (COS) cells</td>
<td></td>
</tr>
<tr>
<td>Mammalian cells</td>
<td>Baby Hamster Kidney (BHK) cells</td>
<td></td>
</tr>
<tr>
<td>Transgenic organisms</td>
<td>Plants</td>
<td>(Gelvin, 1998; Janne et al., 1998)</td>
</tr>
<tr>
<td>Transgenic organisms</td>
<td>Rabbit</td>
<td></td>
</tr>
<tr>
<td>Transgenic organisms</td>
<td>Cattle</td>
<td></td>
</tr>
</tbody>
</table>

In contrast to prokaryotic hosts, eukaryotic expression systems are capable of performing post-translational modifications. Such modifications, like glycosylations, which also can differ a lot between different eukaryotic hosts, are sometimes needed to retain the biological activity of a protein.

Yeast is, due to its relative simplicity, often the first choice when looking for a eukaryotic system for recombinant protein production. Several strains are used including the baker’s yeast *Saccharomyces cerevisiae* (Gellissen et al., 1992) and *Pichia pastoris* (Cregg et al., 2000). Among mammalian cells, COS cells and Chinese hamster ovary (CHO) cells are normally used for production of recombinant therapeutic proteins. Expression vectors are often virally based. Mammalian cell systems offer the most human cell-like conditions which is required for the production of some human proteins, but they have drawbacks in cost, degree of difficulty and sometimes the need of biological supplements which could cause viral contaminations. Insect cells or fungal cells are commonly used as good compromises of an advanced eukaryotic system, but with less drawbacks than mammalian cells.
The development of cell-free strategies based on purified enzymes is an expanding field in biotechnology. Two approaches have guided the efforts to achieve cell-free translation. The first, developed over the past decade, is based on crude cell extract, but has the drawbacks of rapid depletion of energy charge and degradation of protein products or templates by proteases and nucleases (Shimizu et al., 2001). These problems have been partly overcome by using a continuous-flow system (Kim and Choi, 1996). The second approach attempts to reconstitute protein synthesis from purified components of the translation machinery. More than 100 different molecules participate in prokaryotic and eukaryotic translation, many of which have been individually purified for biochemical studies of their functions and structures. Several systems have been described in the literature using different kinds of component combinations consisting of both recombinant and artificial molecules together with native ones (Shimizu et al., 2001). In the future, in vitro translational systems will most probably become more and more widely used as the systems are further developed and refined. The high degree of controllability is of high interest for the pharmaceutical industry. Also the possibilities to optimize all the involved processes to give high yields and pure products are promising.

As can be understood, there are a vast number of parameters that has to be considered before deciding on a particular system for the production of a given protein. In proteomics applications, where throughput is a key feature, E. coli is probably the most promising host cell to create general and robust expression systems without extensive cost or labor-intensive handling.

3.2. Affinity tags

There is a great interest in developing methods for fast and convenient purification of proteins. A powerful technique made possible by the introduction of genetic engineering is to purify the target protein by the use of a genetically fused affinity fusion partner. Such fusion proteins can often be purified to near homogeneity from crude biological mixtures by a single affinity chromatography step.

To date, a large number of different gene fusion systems, involving fusion partners have been described, using different types of interactions such as enzyme-substrate, polyhistidines-metal ions, bacterial receptor-serum protein and antibody-antigen. (Uhlén et al., 1992) Some of the most commonly used systems are listed in Table 3. Combinatorial protein chemistry has also opened up new possibilities to screen for affinity ligands with novel binding specificities.
Staphylococcus aureus protein A (SPA) and the streptococcal protein G (SPG) are both bacterial receptors present on the surface of the gram-positive bacteria and can bind to the Fc part of IgG molecules and serum albumin, respectively. Several common affinity fusion partners of varying sizes are derived from domains of these two bacterial proteins, for example the Z domain derived from the B domain of SPA and the albumin binding protein (ABP) derived from the albumin binding region of SPG. Both SPA- and SPG-derivatives are easily produced in bacterial systems such as E. coli, and they are known also to increase the yield and the overall solubility of the produced fusion protein (Ståhl et al., 1999). ABP has also shown immunopotentiating effects when fused to an immunized antigen (Sjölander et al., 1997; Libon et al., 1999). Both SPG and SPA fusion proteins are most conveniently eluted from affinity columns using low pH.

Table 3. Commonly used affinity fusion systems.

<table>
<thead>
<tr>
<th>Fusion partner</th>
<th>Size (kDa)</th>
<th>Ligand</th>
<th>Elution</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein A</td>
<td>31</td>
<td>hIgG</td>
<td>Low pH</td>
<td>(Nilsson and Abrahmsén, 1990)</td>
</tr>
<tr>
<td>Z</td>
<td>7</td>
<td>hIgG</td>
<td>Low pH</td>
<td>(Nilsson et al., 1987)</td>
</tr>
<tr>
<td>ABP, ABD</td>
<td>5-25</td>
<td>HSA</td>
<td>Low pH</td>
<td>(Nygren et al., 1988)</td>
</tr>
<tr>
<td>His&lt;sub&gt;6&lt;/sub&gt;</td>
<td>1</td>
<td>Ni&lt;sup&gt;2+&lt;/sup&gt;, Co&lt;sup&gt;2+&lt;/sup&gt;, Me&lt;sup&gt;2+&lt;/sup&gt;-chelators</td>
<td>Low pH/ imidazole</td>
<td>(Porath et al., 1975; Porath, 1992)</td>
</tr>
<tr>
<td>GST</td>
<td>26</td>
<td>Glutathione</td>
<td>Reduced glutathione</td>
<td>(Smith and Johnson, 1988)</td>
</tr>
<tr>
<td>FLAG</td>
<td>1</td>
<td>mAb M1, M2</td>
<td>EDTA/low pH/ FLAG peptide</td>
<td>(Hopp, 1988)</td>
</tr>
<tr>
<td>MBP</td>
<td>40</td>
<td>Amylose</td>
<td>Maltose</td>
<td>(di Guan et al., 1988)</td>
</tr>
</tbody>
</table>

A very robust and general affinity method is the immobilized metal-ion affinity chromatography (IMAC), which is based on the interaction between positively charged metal-ions and negatively charged amino acids (Porath et al., 1975; Porath, 1992). Metal-ions such as Co<sup>2+</sup> or Ni<sup>2+</sup> are immobilized on a resin and a histidine-rich sequence is added genetically to the produced protein. The tag could be placed either N- or C-terminal, the system can be used under either denaturing
or native conditions and elution is achieved by moderately low pH or imidazole. These features have made IMAC one of the most easily accessible systems for affinity purification. The major drawback is that it is not based on a biospecific interaction, which makes it less specific. Native proteins with many histidines are easily co-purified to a certain extent.

Glutathione S-Transferases are a family of enzymes that can transfer sulfur from glutathione to substances such as nitro and halogenated compounds, leading to their detoxication (Mehler, 1993). Many mammalian GSTs can be purified by affinity chromatography using the immobilized cofactor glutathione followed by competitive elution with reduced glutathione (Simons and Vander Jagt, 1981). The elution method is also a possible complication as the use of reduced glutathione may affect target proteins containing disulfides (Sassenfeld, 1990).

Combinatorial protein chemistry has opened up new possibilities to make tailor-made affinity ligands (Nygren and Uhlén, 1997). One such example is the affibodies (Nord et al., 1995 and 1997), where the Z domain, derived from protein A, and which normally binds to the Fc part of IgG, is used as a scaffold. Thirteen amino acids on the binding surface out of a total of 58 amino acids are randomized to give the molecules new binding specificities. Novel binders to different targets can then be selected from the library by phage display or other in vitro selection methods (Nord et al., 1995, 1997 and 2000; Hansson et al., 1999; Gunneriusson et al., 1999).

4. Antibodies as tools in proteomics

The binding, reversible or irreversible, of one molecule to another is a key moment in almost all biological processes. For one molecule to be able to interact with another it first has to find its partner and then attach to it somehow to be able to interact. In the field of functional proteomics there is a great need for molecules with specific and strong affinities. These binders could be used in many applications described earlier, such as localization studies and purification methods. The dream would be a complete library with binders to all members of all proteomes to perform high-throughput immunohistochemical studies. Specific antisera towards different proteins can also be used for isolation of native protein complexes from its native environment, followed by identification by mass spectrometry (Shevchenko et al., 1997). Specific binders could also be of great use in many therapeutic applications, to guide a drug or a toxin to an exact location in the organism.
In nature there exists a system which is capable of generating binders to almost anything. This is the immune system and its highly variable antibody molecules. Due to randomized exon shuffling, the human body is capable of producing approximately $10^9$ antibodies with different binding specificities and affinities. Via immunizations of animals or by screening of monoclonal antibody libraries, antibodies specific to a target protein can be achieved. Protein for immunization can be obtained either by chemical synthesis of peptides or by recombinant expression of a part or the complete protein. Protein could also be purified from its native source, but this type of procedure is usually very labor-intensive and requires optimization for every specific protein.

4.1. Recombinant antibodies

Antibodies are quite difficult molecules to produce by recombinant means. They have a large and complex structure held together by a couple of disulfide bonds. These features make antibodies impossible to produce in prokaryotic systems. Several smaller versions of antibodies have been developed, such as the Fab fragment and the single-chain Fv (scFv) fragment. A Fab fragment consists of one of the light chains of an antibody plus the corresponding part of one of the heavy chains held together by a disulfide bond. Cleavage of an IgG molecule results in one Fc fragment and two Fab fragments. The scFv molecule consists of the variable parts of the light and the heavy chains, respectively, held together by a flexible peptide linker (Figure 2).

ScFvs are easier to produce than Fabs, but when the intention is to eventually generate a whole antibody molecule, a Fab is a better starting point. By amplification of total mRNA from certain cell types, all different variants of antibodies can be recovered together and cloned into a vector. This mixture of maybe billions of variants is called a library and can be displayed on bacteriophage (McCafferty et al., 1990), lytic phage (Forrer and Jaussi, 1998), bacteria (Daugherty et al., 1998), yeast (Kieke et al., 1997), ribosomes (Hanes and Pluckthun, 1997) or may be linked to DNA in droplets of emulsion (Tawfik and Griffiths, 1998). By the use of such in vitro selection systems, special affinity qualities can be screened for. To get affinities that can be compared to those of nature-made antibodies, one usually has to use methods of affinity maturation. In this way the processes of the immune system can be mimicked to obtain the desired binding molecules without having to use live animals. Furthermore, since no immunization steps are required, comprehensive phage-antibody libraries
permit targeting of antigens in vitro, which are known to be toxic and/or possess low antigenicity in vivo, such as self-antigens or antigens which are highly homologous between species (Krebs et al., 2001).

Figure 2. Schematic views of a human IgG molecule, a Fab fragment and a single-chain Fv fragment, respectively.

Technology platforms for high-throughput generation of recombinant antibodies are presently being developed by several companies and research groups. A library called Human Combinatorial Antibody Library (HuCAL®) based on phage display of scFvs have been optimized for high-throughput generation and targeted engineering of human antibodies (Krebs et al., 2001). So far, the library was reported to have been charged with 117 targets for which all, specific scFv antibodies could be identified. This system is not only a rich source of human antibodies for various therapeutical applications, but also a valuable tool in large-scale functional proteomics efforts.

4.2. Antibody arrays

In order to monitor thousands of interactions, thousands of different antibodies would need to be screened. By arraying the antibodies and performing parallel screens using the same antibody array on different tissues it should be possible to identify antibodies that bind to differentially expressed proteins, even when the
individual identities of the proteins are not known (Holt et al., 2000). For arrays of tens of thousands of members, the use of polyclonal antibodies derived by animal immunization will not be practical. Recombinant antibodies, however, are ideally suited for the creation of such arrays and are generally expressed at sufficiently high concentrations in the bacterial supernatant for direct use (Holt et al., 2000).

Like DNA arrays, antibody arrays can take many forms. From the spatial patterning of just a few molecules on a solid support to a high-density microarray, the latter having the advantage of requiring very small sample volumes. The array can be arranged on a variety of supports including glass, nitrocellulose, polyvinylidene fluoride membranes (PVDF), polystyrene or other plastic materials (Cheung et al., 1999). Detection methods of common use include enzyme-linked chemoluminiscence, radiolabelling, mass spectroscopy and surface-plasmon resonance (Cheung et al., 1999). To immobilize the antibodies on the support biotinylated antibodies have been used on a streptavidin-coated surface (Silzel et al., 1998). The antigen can also be coated onto a solid support and the antibodies be allowed to bind to it instead. Microarrays are preferably spotted onto glass as it has greater durability and lower intrinsic fluorescence. The spotting can for example be performed with a pressure-controlled capillary system onto an N-hydroxysuccinimide (NHS)-activated glass slide (Mendoza et al., 1999).

An alternative to using pure antibody molecules or bacterial supernatants to create an array is to directly spot bacteria containing cloned antibody genes onto a filter (de Wildt et al., 2000). Since the antibody molecules are excreted, the filter with bacteria can be used to make several stamps on other filters coated with different antigens to be analyzed. Although the density of these arrays is limited by the irregularity of bacterial growth, parallel screens of tens of thousands of different antibodies against different complex antigens can be performed (de Wildt et al., 2000).
PRESENT INVESTIGATION

The work presented in this thesis concerns different methods of analysis of cDNA-encoded proteins on a functional basis. The large-scale sequencing efforts that have been performed during the last decade have generated an abundance of partial cDNA sequences accessible in the public databases. A method to retrieve full-length cDNA clones from a partially known sequence is described in paper I. Functional characterization of cDNAs and their corresponding proteins require general and robust methods for protein expression and purification. In papers II-V, the development of several high-throughput expression systems enabling proteomic approaches to gene characterization, are described. The main approach is to use the produced cDNA-encoded proteins to generate antibodies specific for the target protein to be used in immunolocalization studies. Characterizations of the human \textit{APC1} and \textit{TEKT1} from \textit{Mus musculus} are described in papers VI and VII respectively, based partly on the methods described in the first five papers.

5. Recovery of upstream cDNA sequences (I)

The amount of partial cDNA sequences, i.e. ESTs, reported to the public databases is enormously large and statistically they cover all genes expressed in several organisms. A lot of information could potentially be gained from this vast pool of small pieces of sequence information. This has created a need for methods to gain more complete coding sequences, corresponding to full-coding mRNA transcripts. Earlier, this has been done by hybridization screening of cDNA libraries using probes representing parts of the known sequence. This method works fine if the full-length sequence is present in the existing library, but in many cases this is not so. Several PCR-based methods have been developed, including rapid amplification of cDNA ends (RACE) (Frohman et al., 1988), anchored PCR (Loh et al., 1989) and one-sided PCR (Ohara et al., 1989). An alternative method that uses capture of the 5' ends via biotinylation has also been described (Carninci et al., 1997). Although these methods are working well in some applications, yielding a full-coding sequence of a transcript can still be a difficult task. To determine an unknown DNA sequence adjacent to known sequences, a method has been described which uses biotin capture, low-stringent arbitrary priming and extension, and direct sequencing without subcloning (Nguyen et al., 1998; Sterky et al., 1998). Paper I describes a similar method, which instead uses RNA as starting template, for recovery of unknown cDNA sequences.
**Figure 3.** The basic concept of the method for recovery of upstream cDNA sequences.

The concept is shown in figure 3, above. A biotinylated primer (T1b) specifically annealing within the known region of the target gene is used for reverse transcription of poly(A)+ RNA yielding a single-stranded cDNA. The cDNA is then immobilized on streptavidin-coated paramagnetic beads. Four primers (A-D) with a common 5’-handle and varying 3’-regions are used for low-stringency priming of the single-stranded template to provide a starting point for the second cDNA strand. The resulting double-stranded cDNA can then be amplified using an upstream primer (E) specific to the common sequence introduced by the extension.
primers (A-D) and a downstream nested primer (T2) which anneal specifically in the known region. The PCR product can then be used directly as template for DNA sequencing with a third, target gene-specific, primer, T3. The arbitrary primers are general in the concept and can be used for any gene for which full-coding sequence is wanted, whereas primers T1b, T2 and T3 are specific for the particular gene.

In paper I the concept was evaluated for the cDNA clone TSG118, a murine nucleolar spermatogenesis-related protein (Larsson et al., 1999), that had earlier failed to give the full-length sequence by conventional library screening. Three primers, T1b, T2 and T3 were designed based on the known region of TSG118. The method was developed and applied as described above using primer B as arbitrary primer in the synthesis of the second cDNA strand. This resulted in two PCR bands with the approximate sizes of 350 and 500 bp, respectively, which were isolated on an agarose gel and used as templates in a semi-nested PCR with primers E and T3. After the reamplification, the PCR products were used directly as templates in cycle sequencing with primer T3. Both sequences extended into the previously known region of TSG118 and had a common 5' sequence of 100 bp containing a methionine codon and upstream of that an in-frame stop codon. Following the assumed start codon, the two fragments have a 24 bp stretch in common after which a 72 bp region is missing from the shorter one. This missing sequence could potentially represent an exon of the TSG118 gene that displays differential splicing, a speculation that still remains to be fully confirmed by the isolation of the two different cDNA clones.

In conclusion, the described method for recovery of unknown upstream sequences based on known regions was shown to be able to produce the complete coding sequence of the TSG118 gene. The method was also able to detect what could potentially be splice variants of this gene, something that would not be found when using methods that include subcloning of the recovered cDNA. This method should thus represent a complement to previously described RACE-based methods for recovery of upstream cDNA sequences.

6. Development of high-throughput expression systems (II-V)

In functional proteomics research, the recombinant production of proteins is fundamental for many applications. The expression systems presented in this thesis have been focused to create general tools for localization studies. A convenient method is to generate specific affinity reagents, for example antibodies, for use in
various immunolocalization analyses. The generation of antibodies has conventionally been performed by immunizing animals with a protein, either the complete molecule or a smaller portion of it. Large efforts are being invested into research to find other ways to generate affinity reagents by \textit{in vitro} selection methods, but still none is as effective as the immune system of a living organism.

However, for immunizations and other applications to generate binding molecules, there is a need for protein material, which can be produced in three ways. First, peptides representing parts of the protein can be synthesized chemically, second, native protein can be extracted from cells and tissues and finally, recombinant methods can be used to produce a protein, full-length or partly, in a host cell. In papers II-V, the development of general bacterial expression systems for recombinant production of cDNA-encoded proteins aimed at the generation of antibodies for large-scale localization studies are presented.

6.1. \textbf{A high-throughput protein expression system}

A first approach towards a high-throughput expression system to generate antibodies, specific for cDNA-encoded proteins, had been performed earlier (Larsson et al., 1996). The concept included two vectors, expressing the cDNA in \textit{E. coli} in parallel as fusions to a staphylococcal protein A (ZZ) tag and an albumin binding protein (ABP) tag (residues 146-266) from streptococcal protein G, respectively. The ZZ fusion protein was purified by affinity chromatography on IgG-Sepharose columns and was used for immunization. The ABP fusion protein was affinity purified on human serum albumin (HSA) columns and used in an immobilized form for affinity enrichment of antibodies specific for the cDNA-encoded protein from the generated antiserum.

Since it was later found that the ABP tag had immunostimulating carrier properties (Sjölander et al., 1997; Libon et al., 1999), it was of interest to include this tag in the fusion proteins to be used for immunizations. The presented dual system (Larsson et al., 1996) mentioned above could however not be reversed since the ZZ fusion protein would bind to all IgG antibodies through its IgG-Fc binding capacity (Nilsson et al., 1987; Ståhl et al., 1999), if used as a fusion protein for affinity capture. In paper II a second concept was presented employing a single vector system where the cDNA was expressed with an N-terminal dual affinity tag consisting of a His\textsubscript{6} tag in frame with the ABP tag. The His\textsubscript{6} tag provides a mean for streamlined purification under denaturing conditions direct from total cell culture. The reasons for including the ABP tag were first to obtain the
immunostimulating effect mentioned above, but also (i) the possibility to perform expression screening using affinity-blotting with biotinylated human serum albumin (HSA) (ii) beneficial production properties such as high expression levels (Murby et al., 1995; Ståhl and Nygren, 1997), and (iii) promotion of solubility (Murby et al., 1994; Jonasson et al., 1998). The ABP tag can also be used as an optional tag for affinity recovery on HSA-Sepharose under native conditions (Larsson et al., 1996).

**Figure 4.** The basic concept for the described high-throughput expression system.
The basic concept is outlined in figure 4, above. Selected cDNA clones are PCR amplified from phage stocks or bacterial colonies using *Pfu* polymerase, with a biotinylated downstream primer and a frame-specific upstream primer. The biotinylated PCR products are then inserted into the constructed expression vector, pAff8c, by a solid-phase subcloning procedure. To minimize internal restriction of the cDNA clones, restriction enzymes with GC-rich eight-nucleotide recognition sites, *Not*I and *Ascl*, are used. The sites are introduced in the PCR fragments by the flanking regions of the primers, respectively. The restricted PCR fragments are ligated into the previously digested pAff8c and the constructs are transformed to the *E. coli* expression host. The transformed colonies are PCR screened to verify insertions. Correct clones are used for expression of fusion proteins followed by affinity recovery by immobilized metal-ion affinity chromatography (IMAC) (Porath et al., 1975; Porath, 1992; Hochuli, 1988). The purified fusion proteins are used for immunization and the resulting antisera are affinity enriched on the corresponding fusion protein, immobilized on a solid support. The highly specific antisera obtained by this procedure could then be used for immunological analyses, e.g. Western blot and immunolocalization. Furthermore, the antisera could potentially be used as affinity ligands for isolation of native protein or protein complexes.

To evaluate the system for expression of mammalian genes, 55 cDNA clones from a cDNA library constructed by a differential cDNA library screening method (Höög, 1991; Starborg et al., 1992; Yuan et al., 1995), were used. The cDNA clones had inserts with sizes ranging from 500 to 1600 bp with open reading frames (ORFs) quite evenly distributed between the three alternative frames. All of the 55 clones were successfully subcloned into pAff8c with 100% efficiency using the solid-phase subcloning procedure mentioned above. Expression of fusion protein was obtained for 49 of the 55 clones with expression levels ranging from 15 to 80 mg/l. This means that this general scheme yielded enough protein for immunization from 50-100 ml shake flask cultures for 90% of the cDNA clones.

For a more detailed evaluation of the complete concept, a subset of five cDNA-clones was selected of which four showed homologies to genes whose proteins had known subcellular localizations. The fifth clone did not display any strong homology at all and was thus selected as a gene with unknown localization. To study the expression efficiency, affinity blotting with biotinylated HSA was performed on whole-cell extracts from the cultivations. Strong bands were observed at harvest while nothing was stained before induction demonstrating tight regulation of the expression system. Furthermore, the only natively biotinylated
protein in *E. coli*, biotin carboxy carrier protein (BCCP) (Fall et al., 1976) was detected in each sample, thus functioning as an internal standard in the blotting procedure.

**Figure 5.** Tissue screening by Western blot analysis for the five clones, 8, 18, 29, 30 and 54. Molecular sizes in kDa are indicated by the right margin.

The proteins from the five selected clones were used to immunize rabbits and the antisera were enriched using the same fusion protein in immobilized form as affinity ligand. Subsequently, the enriched antisera were used in immunological analyses, e.g. Western blotting and immunofluorescence microscopy. Western blots were performed on protein extracts prepared from testis, epididymis, proliferating Swiss-3T3 cells, pachytene spermatocytes and round haploid cells. The five antisera selectively reacted with unique protein bands with molecular masses of 105, 100, 75, 200 and 70 kDa, respectively (Figure 5). The expression patterns were quite different for the five clones, where some were expressed in several cell types and some predominantly in one. Tissue screening could in itself give valuable information about protein expression, but also important guidance in
the selection of which cell types to study in more detail in and immunohistology analyses.

Three of the five clones, all with homologies to proteins with known subcellular localization, gave highly specific staining patterns in the indirect immunofluorescence microscopy studies. They also showed a specific overlap when double staining with antibodies against those subcellular structures, the sperm tail, actin filaments and cell surface, was performed. The other two antisera, of which one had no known homology, showed signals too weak to be interpreted in the analyzed cell types.

In conclusion, a system for generation of antibodies, highly specific to cDNA-encoded proteins, to be used in immunolocalization studies, has been developed and evaluated. The system is built on robust unit operations designed for the possibility to be automated. The expression efficiency of mammalian proteins in a bacterial host was over 90% and informative localization results were obtained by Western blotting and immunofluorescence microscopy.

6.2. A high-stringency dual concept

Based on the results obtained from the previously described high-throughput expression system and the one described by Larsson et al, 1996, a third generation expression system was developed. The single-vector strategy employed in the system described in paper II, meant less manual labwork to be performed than with the dual system described earlier (Larsson et al., 1996). However, the single-vector system also resulted in the co-enrichment of antibodies specific to the ABP tag. Validation blots, i.e. quality testing in Western blot of the enriched antibodies reactivity towards the fusion protein used for immunization, would be inconclusive from the single-vector approach, since the affinity tag with the ABP portion would be present in all fusion proteins. This was found to be a significant drawback since such quality testing was found to be highly informative before using the antibodies in extensive functional annotation studies.

Therefore, in paper III, we decided to re-evaluate a dual expression system, keeping the positive results from the single-vector approach (Figure 6). Here the cDNA is expressed in parallel in pAff8c (constructed in paper II) in fusion with His6-ABP for the immunizations, and in a new vector, pAff10c, in fusion with an N-terminal His6 tag alone for affinity enrichment of antibodies. Since the His6 tag is so small, the risk of enrichment of antibodies reactive to the affinity tag is
significantly reduced when using these fusion proteins for the enrichment procedure.

**Figure 6.** The basic concept for the presented dual expression system.

To evaluate the system, the same five cDNAs that were studied in some more detail, and also immunized, in paper II, were subcloned into the new vector pAff10c, and expressed as fusions with a His$_6$ tag. The fusion proteins were recovered using a similar IMAC procedure as in paper II. The expression levels were in the same order of magnitude as for pAff8c, ranging from 10 to 55 mg/l. Antibodies from the antisera generated earlier (paper II), were affinity enriched with the His$_6$-cDNA fusion proteins immobilized as affinity ligands onto nitrocellulose filters. The enriched antibodies were subsequently tested in a
validation blot towards the corresponding His\textsubscript{6}-cDNA fusion proteins. All five antibody preparations gave specific stainings of the corresponding fusion proteins, but not to another unrelated His\textsubscript{6}-tagged control protein, indicating that the antibodies were specific for the cDNA-encoded part.

To test the antibodies in a more relevant biological setting, the enriched antibodies were used in Western blots against the same cell extracts as described in paper II. Three of the antisera, 8, 18 and 30, stained specific and unique protein bands (Figure 7) and interestingly, these reactivities differed significantly from those observed earlier (paper II). The two clones, 29 and 54, that did not stain any specific bands in this study had earlier stained a protein of approximately 70 kDa in size, in Western blot analysis of the same cells and tissues. For the other three the Western blot analysis gave similar results as before, except that for clone 18, a new band of approximately 150 kDa appeared in epididymis and another band for clone 8, of approx. 105 kDa, disappeared from the epididymis lane (Figure 7).

![Figure 7](image)

**Figure 7.** Tissue screen by Western blot analysis for clones 8, 18, 29, 30 and 54, respectively with more stringently enriched antibodies. Molecular sizes in kDa are indicated by the right margin.
In paper II, immunofluorescence studies failed for clones 29 and 54, both on tissue sections and on purified cells, while the three other clones, 8, 18 and 30, generated antibodies that gave highly specific and informative localization information. It would thus be tempting to speculate that the reactivities found earlier for clone 29 and 54 were indeed false positives. Then the present dual expression concept represents a more stringent strategy, with a stronger capacity to avoid falsely positive results. The other discrepancies for clone 8 and 18 are more difficult to interpret, and would need further experimental studies to be confirmed, but could very well be of biological significance.

In conclusion, this dual system has significantly improved the high-throughput expression system described in paper II. The system would have the potential of being highly suitable in various immunoproteomics studies.

6.3. A novel affinity tag allowing protein A-based recovery

In paper IV, an expression vector system has been developed taking advantage of a novel Staphylococcus aureus protein A (SPA)-binding affinity tag, a Z-affibody denoted Z_{SPA-1} (Eklund et al., 2002), which enables straightforward affinity blotting procedures and efficient affinity recovery. The Z_{SPA-1} affibody was previously selected by phage display from a library constructed by combinatorial mutagenesis of a protein domain of SPA (Nord et al., 1995, 1997 and 2000; Gunneriusson et al., 1999). The tag enables affinity recovery on protein A-Sepharose which is a well-documented, robust and readily available chromatography medium. An E. coli expression vector was constructed for intracellular production under the control of the T7 promoter system (Studier et al., 1990), with an N-terminal dual affinity tag consisting of a His_{6} tag in frame with Z_{SPA-1} (Figure 8).

Earlier, an affinity purification method taking advantage of the interaction between protein A, used as affinity tag, and the Fc part of IgG molecules in the form of IgG-Sepharose has been used (Ståhl et al., 1999). A drawback of this system is that the IgG molecule, employed as immobilized ligand on the chromatography medium, is relatively fragile and frequently associated with leakage of antibody subdomains in the eluate. Furthermore, commercially available IgG resins consist of polyclonal immunoglobulin preparations derived from human donors, which is a major drawback if the products are intended for therapeutic use due to the risk of virus contamination.
The affibodies are based on a 58 amino acid three-helix bundle protein scaffold, where 13 amino acids on the surface have been randomized to give novel binding capacities. The absence of disulfide bonds, high inherent solubility and proteolytic stability, should make the affibodies suitable as fusion partners for expression in bacterial hosts.

To evaluate the system, the same five cDNA clones that were selected and produced in pAff8c and pAff10c (paper II and III), were subcloned and expressed in pAff11c. To compare purification efficiencies, the fusion proteins were purified both by IMAC using the His6 tag and on protein A-Sepharose using the ZSPA-1 tag. It was found that the protein A-based recovery resulted in the highest degree of purity. Two purification protocols were used for the protein A-Sepharose purification due to the variable solubility properties of the produced fusion protein. The highly insoluble products that were produced mainly as inclusion bodies, had to be denatured with guanidinium hydrochloride and then renatured prior to the purification procedure. Thus, it was of interest to investigate if an alternative more general recovery concept, independent of the target protein solubility, could be developed. The total cell pellet was denatured containing both soluble and insoluble products, and subsequently partly renatured to different concentrations of Gua. The affinity recovery procedure was then performed with a corresponding

Figure 8. Schematic picture of the expression vector pAff11c.
amount of Gua in the buffers. It was shown that protein could efficiently be recovered in the presence of as much as 0.5 M Gua. Protein could also be recovered in the presence of 1 M Gua, but with significant losses.

Two convenient blotting procedures were successfully developed for screening of expression efficiencies, one two-step procedure based on commercially available reagents and one single-step procedure using in-house produced reagents. The first affinity blotting procedure used recombinantly produced protein A, thus binding to ZSPA-1, followed by peroxidase-anti-peroxidase antibodies, and subsequent development with peroxidase substrate. The single-step method employed a protein A-alkaline phosphatase fusion protein, which could directly upon binding be developed by addition of phosphatase substrate.

In this paper we have described an expression system allowing protein A-based affinity recovery of non-immunoglobulin products. The system, which employs a tailor-made protein A-based affinity tag for affinity purification an protein A-Sepharose and with convenient blotting procedures for expression screening, constitutes an expression system which should have a potential in high-throughput proteomics efforts.

6.4. Single vector three-frame expression systems

In paper V, an effort to create expression vectors which would allow expression of an inserted gene fragment in all three reading frames simultaneously, is presented. The vectors have one promoter, but three separate ribosome-binding sites (RBS) and ATG start codons, each in frame with an N-terminal affinity tag to allow efficient recovery of the produced fusion proteins. The idea is to combine cloning and expression vectors into one construct for the production of cDNA-encoded proteins for functional analysis studies. In order to generate proteins, a typical first step is to create a cDNA library in a cloning vehicle, from which open reading frames can be established and interesting clones be selected for expression. These are then subcloned into a second vector to enable protein expression in a suitable host. The system described in paper V was designed to combine these two functions, which is made possible by the ability to express an insert in all three frames.
Figure 9. The expression vector p3xHis (A) with ZZ inserted in three frames (B).

Both vectors presented in this paper are designed for intracellular production with tightly controlled transcription by the phage T7 promoter system (Studier et al., 1990). In the first vector, p3xHis (Figure 9a), three identical hexahistidyl tags are used as affinity tags whereas in the second vector, pAff9 (Figure 10a), three different tags are used. In pAff9 expression in the first frame will generate a fusion tag consisting of a 46 amino acid residue albumin binding domain (ABD) (Nygren et al., 1988; Ståhl et al., 1999) derived from streptococcal protein G (SPG). Expression in the second frame will generate a fusion tag consisting of a 58 amino acid residue Z domain derived from staphylococcal protein A (SPA) (Nilsson et al., 1987), and the third frame will generate a hexahistidyl tag. The gene encoding the Z tag was redesigned and synthetically made by solid-phase assembly of oligonucleotides to avoid stop codons and dimeric basic amino acids in the two non-coding frames. Both expression vectors have been evaluated through the expression of model gene products, inserted for expression in three reading frames.
Figure 10. The expression vector pAff9 (A) with SCP1 inserted in three different frames (B).

For p3xHis, the divalent protein domain ZZ was selected to serve as model protein and was thus inserted in three frames into the vector (Figure 9b). The three constructs were expressed in E. coli cultivations and fusion proteins were recovered by IMAC. All three fusion proteins could be readily detected, with clearly visible size differences, which demonstrated that all three start sites were functional. Expression of the first frame was significantly favored over the second and the third start sites. This might be due to easier access to the first RBS for the ribosomes. It was furthermore unexpectedly observed that the hydrophobic isoleucine sequence, resulting when the His6 tag encoded by CAT codons was translated out of frame, seemed to give hydrophobic interaction effects that made the recovery procedure less efficient.

The success of creating a vector capable of expressing all three frames simultaneously, encouraged the design of a second generation vector denoted pAff9. To be able to separate the three produced fusion proteins, three different affinity tags were used, and the “isoleucine effect” would be circumvented by using the other possible His codon to encode the His6 tag. The cDNA encoding the
Functional proteomics: Generation and analysis of cDNA-encoded proteins

Synaptonemal complex protein 1 (SCP1) from Mus musculus was employed as model protein, and 400 bp of the N-terminal was subcloned in three different frames in pAff9 (Figure 10b). The constructs were expressed in E. coli and the harvested cells from each cultivation were divided in three parts, each to be affinity purified on HSA-sepharose, IgG-sepharose and by IMAC, respectively. Whole cell extracts were analyzed by SDS-PAGE and affinity blotted with biotinylated HSA to detect the ABD fusion proteins and PAP-stained to detect the Z-fusion proteins, respectively. Protein bands of expected sizes were detected for both methods, but in harmony with the results from p3xHis, a significantly higher expression was observed for the first start generating ABD fusion proteins, than for the second start. The purified proteins with the correct inserts were also analyzed by SDS-PAGE and Western blotted with antibodies specific to the N-terminal part of SCP1 (Liu et al., 1996). The fusion proteins ABD-SCP1 and Z-SCP1 were readily detected, but His₆-SCP1 could not be detected in our assays. This might be due to the long stretch between the end of the mRNA and the third RBS making it impossible for the ribosomes to find this RBS. Since transcription and translation are coupled events in prokaryotes, the end of the transcript can bind to a ribosome as soon as it starts to emerge, and thus the later start sites will be blocked by ribosomes already at work translating from the first start.

Another unexpected result was the detection of ABD-SCP1 in the correct frame in cultivations of constructs that should not be able to express this fusion protein. The reason for this is not obvious and could only be explained by a transcriptional frameshift, or perhaps more likely, a translational frameshift event occurring between the ABD tag and the SCP1 insert. Since no proteins with both the ABD and the Z tags functional were detected in the affinity blotting assays, the frameshift is likely to occur in close proximity to the His₆ sequence. But as the effect was not observed for the p3xHis vector where another His codon was used, it is tempting to speculate that the slippage occurs over the repetitive sequence generated by the six CAC codons.

Although these single vector three-frame expression systems are not fully evaluated and optimized, still the principle to create a functional single vector three-frame expression system has been demonstrated. To improve the system further, an alternative coding sequence for the His₆ tag, using both CAC and CAT, might help. Different Shine-Dalgarno sequences with varying strength could also be used to make the expression efficiencies from the three start sites more even, i.e. put a weak one first, a neutral in the middle and a strong one last. The use of smaller tags put in size order might also counteract the stronger expression from
the first start. Furthermore, it would be highly interesting to evaluate the vector in an in vitro transcription/translation system. Since the two events could then be separated, the RBSs would be exposed to the ribosomes on more equal terms.

In conclusion, this is the first successful attempt to create a three-frame expression system. Although further investigation is needed, this concept could become valuable in future combined cloning-expression vector systems.

7. Functional analysis based on protein expression (VI-VII)

The methods described in the first five papers have been focused on developing technology platforms for functional analysis of proteins, e.g. expression systems for cDNA-encoded proteins and the generation of antibodies with specific reactivity towards these proteins. In the last two papers the main focus is the functional studies of a protein complex and a spermatogenesis-related protein, with the aid of generated target protein-specific antibodies.

7.1. Characterization of the human \textit{APC1}

In paper VI the human gene \textit{APC1}, encoding the largest subunit of the Anaphase-Promoting Complex or Cyclosome, APC/C, is characterized. The APC/C is a ubiquitin ligase, consisting of 10-12 subunits, which targets cell cycle regulatory proteins for degradation by the proteasome. Thus the protein complex plays a significant role in the regulation of mitosis, a process which often is hampered in cancer cells. In paper VI the human \textit{APC1} cDNA is analyzed and the expression of APC/C components is investigated. The full-length human \textit{APC1} cDNA was sequenced and using the program est_genome (Mott, 1997) 42 exons could be predicted covering most of the cDNA sequence. The human \textit{APC1} gene was also mapped to chromosome 2, q12-q14.1.

To investigate whether APC/C components are expressed at the same levels in normal human tissue and human cancer cells a Northern blot analysis was performed. Probes from \textit{hsAPC1}, \textit{hsAPC2}, \textit{hsAPC6} and \textit{hsAPC8} were hybridized to filters containing mRNA from different human tissues or different cancer cell lines. All of the tested APC/C components seem to be ubiquitously expressed in most tissues and the cancer cell lines analyzed. If one of the components was more strongly expressed in a cell, the amounts of the others also increased, supporting the idea that the APC/C components function as a protein complex in all cell types. The only deviation to this was the RNA expression of \textit{APC6} that was somewhat
increased in a few tissues and cell types, indicating a special function for this RNA in these cell lines.

To complement the RNA blot data, the expression of all known human APC/C components was studied through EST data from the public databases using BLAST 2.0. The data supports the Northern blot results and also indicates that the APC6 is expressed in higher levels than the other components in colon, liver/spleen and white blood cells, again suggesting a special function for this protein in these cell types.

Protein expression of APC/C components was also investigated. A region from the C-terminal part of hsAPC1 was subcloned into pAff8c (constructed in paper II), expressed in E. coli and affinity purified. The purified protein was immunized in rabbit and polyclonal antibodies were enriched from the generated antiserum. Antibodies towards the human Apc4 were generated by immunization of chicken with a synthesized 15 amino acid peptide. Antibodies against the human Apcl and the human Apc4 proteins were used to perform Western blots on extracts from different human cancer cell lines. Both the hsAPC1 and hsAPC4 proteins are strongly expressed in all cancer cell lines tested, and protein levels vary only slightly between the different cell lines, suggesting that the APC/C is needed to some extent not only in normal cells, but also in different cancer cell lines and that all subunits are required. It has been suggested that the APC/C may perform functions outside of the cell cycle (Gieffers et al., 1999). In view of the facts that APC/C is expressed in most tissues, even though the majority of cells in most tissues are differentiated and no longer divide, that all APC/C components seem to be required and that the APC/C is a lot more complex than most ubiquitin ligases, it would not be surprising to find that the APC/C components were involved in ubiquitination also of several non-cell cycle targets.

7.2. Tekt1 – the first mammalian tektin

In paper VII the expression of the Tekt1 protein in mouse testis is characterized by immunological methods, e.g. Western blot and immunolocalization. Tektins are a group of filamentous proteins that are known to associate to the outer doublet microtubules which are some of the structural components of cilia, hair-like structures that protrude from many cell types. There are nine outer doublet tubules consisting of an A and a B variant surrounding an inner core (Gibbons, 1981). Tektin filaments are associated to the A-tubule near its inner junction to the B-tubule where they display a periodicity in localization along the tubule. This
periodical location suggests a possible role as positional markers for the different components connected to the A-microtubule. Tektins are also thought to provide stability and beside their association to cilia and flagella, they are also found in centrioles. Tektins were firstly characterized in sea urchins where three variants, A, B and C, are components of the microtubuli of cilia, flagella and centrioles (Linck et al., 1985; Linck and Stephens, 1987; Steffen and Linck, 1988; Norrander et al., 1996).

Centrioles are closely related to the basal bodies from which cilia and flagella are formed and the substructure of centrioles comprise nine microtubule triplets. Centrosomes, the major site for organization of microtubuli, contain two centrioles located perpendicular to each other. The two centrioles duplicate before mitosis begins, producing two daughter centrosomes which then move to the opposite sides of the cell making up the two poles of the mitotic spindle.

**TEKT1** had earlier been shown to be strongly expressed in mouse testis cells (Norrander et al., 1998). A DNA fragment encoding 337 (89-426) amino acids of the Tekt1 protein was expressed and used for immunization. The generated antiserum was affinity enriched and used in Western blot analysis revealing a strong expression of Tekt1 in testis, seminiferous tubules, pachytene spermatocytes and round haploid cells. Relatively low expression was observed in epididymis and none was detected in ovaries and in Swiss-3T3 cells. This result indicates that Tekt1 is not ubiquitously expressed throughout an organism. Furthermore, the expression in spermatocytes and haploid cells suggest that Tekt1 plays a specific role during spermatogenesis.

Immunolocalization studies were performed to establish the subcellular localization of Tekt1 during spermatogenesis. In frozen sections of testis a specific signal was found close to the lumen of the seminiferous tubule, in cells morphologically identified as spermatids. When isolated germ cells were analyzed, the round haploid cells showed a specific staining to a structure at the outer periphery of the nuclear membrane. The displayed subcellular localization suggested centrosomal staining which was confirmed through double-labeling immunofluorescence microscopy with antiserum against the centrosome specific marker ANA. When analyzing a mixed population of purified testis cells, a strong staining of elongated spermatids was observed. The labeling was not localized in the same distinct way as for the centrosome labeling in round haploid cells, but rather displayed a diffuse staining of the caudal end of the sperm head, again overlapping with the localization of the centrioles and the sperm tail development.
No specific staining was detected in the spermatocytes even though Western blot analysis indicated expression of Tekt1 in these cells. This can either be due to a diffuse distribution of Tekt1 within the cell or that it is sterically hindered from interaction with the antiserum. The mature spermatozoa also lack the labeling of Tekt1 antiserum, which may be due to antigen unavailability, but also an actual down-regulation/depletion of Tekt1 at this stage.

In conclusion, by the use of antibodies generated from bacterially expressed portions of Tekt1, the expression of the proteins could be analyzed by Western blot analyses and immunolocalization studies. The expression displays a temporal and spatial correlation with the development of the sperm tail suggesting that Tekt1 may have a specific role in the spermatogenesis process.

8. Concluding remarks and future perspectives

At present, the field of functional proteomics is developing rapidly and a lot of effort and resources are being invested in proteomics-related projects world-wide. There is no doubt that the near future will bring us several new technologies that will help us to gain knowledge about genes and their functions. In this thesis, several methods that are useful in functional analysis of cDNA-encoded proteins have been described.

First, a method for recovery of cDNAs with complete upstream coding sequences was described and applied to obtain the full-coding sequence for a model gene. The concept combines the benefits of low-stringency arbitrary priming, biotin-capture, specific PCR amplification and direct cycle sequencing with the PCR product as template, and would constitute a complement to existing strategies for recovery of upstream cDNA sequences.

The largest part of the thesis describes the development of several expression systems designed for high-throughput expression and purification of cDNA-encoded proteins aimed to raise antibodies specific for the cDNA-encoded proteins. Antibodies generated by subsequent immunization of the produced proteins could then be used in functional analysis studies, such as tissue screening by Western blotting and immunofluorescence/immunohistology localization studies. The first system employs a single vector strategy in which the cDNAs are produced as fusions to a hexahistidyl tag and an immunopotentiating albumin binding protein tag. The His₆ tag allows purification by IMAC under denaturing conditions, thus allowing parallel handling of numerous proteins regardless of their
solubility. The success rate (>90%) in the production and single-step recovery indeed demonstrate the potential for the system to be used in high-throughput efforts. The high specificity demonstrated by the produced antisera suggests that the presented approach would be a useful strategy to determine cellular and subcellular localization of cDNA-encoded proteins. The second system, which employed a dual vector system, offered a more stringent way for affinity enrichment of the antibodies, than the single vector system. The cDNA-encoded protein was first produced in the same vector as the first system with the immunopotentiating affinity tag and was subsequently immunized in rabbits. It was also produced in the second vector as fusion to a His\textsubscript{6} tag only, and this second protein was used in an immobilized form to enrich the antisera for cDNA-specific antibodies.

An expression system employing a novel affinity tag allowing protein A-based affinity purification was also described. An affibody binding to protein A had earlier been selected and was here introduced to function as affinity tag in a new vector which was evaluated by the expression of five cDNAs. A purification scheme was developed and two convenient affinity blotting methods to examine expression efficiency, prior to purification of the expressed fusion proteins, were described. The presented expression system, employing a tailor-made affinity tag for recovery of expressed fusion proteins on a robust commercially available chromatography medium, would indeed have the potential to become widely used for high-throughput expression.

An effort to create single vector three-frame expression systems was also described. The aim was to create a combined cloning and expression vector for more efficient handling of large amounts of cDNAs. Although the two systems described needs further optimization, the principle of three-frame expression was demonstrated.

Antibodies generated according to the principles described in this thesis were used in two studies for functional analysis of cDNA products. The gene encoding the main component of the Anaphase-Promoting Complex or Cyclosome, \textit{APC1}, and the gene \textit{TEKT1} (encoding mouse tektin C), were subjected to functional characterization. The APC/C components were shown to be expressed in equal amounts in almost all cell types indicating that they are indeed forming a complex, which is critical for survival. The expression of the \textit{TEKT1}-encoded protein Tekt1 in mouse testis was investigated. The localization of Tekt1 in haploid cells during
spermatogenesis indicates a possible involvement of Tekt1 in the sperm tail formation.

In contrast to genome sequencing, the mapping of proteomes cannot be done with one single and straightforward method due to the larger complexity of the proteome. Many of the technologies used today have been developed many years and even decades ago, but the strong efforts put into the proteome projects now brings hope of the breakthrough of new strategies. Also, existing methods need to be modified to suite large-scale global studies. The expression systems described in this thesis could easily be further automated and streamlined to allow real high-throughput expression and in fact, has already been used to express all predicted genes on human chromosome 21. The produced fusion proteins have been immunized and the generated antisera have been tested in tissue screening and immunolocalization studies with high efficiency (Agaton, 2002). However, the bottleneck of this system is clearly the generation of antibodies by immunization, which is limiting both regarding to cost and the ethical aspect of large-scale handling of laboratory animals. The hope lies in the possibility to mimic the immune system by *in vitro* selection methods. Binding molecules could either be selected from libraries of antibody-like molecules such as Fabs or scFvs (Krebs et al., 2001), or a library of novel types of affinity reagents such as affibodies. Selection methods can either be based on classical phage-display procedures or other novel selection principles that are being developed.

If affinity reagents were available for a majority of the members of the human proteome and those of other species, global functional analysis studies would be enabled using different affinity reagent-based techniques. One informative strategy is to create protein arrays for protein expression profiling. Many large-scale projects will most likely be reported in the near future to come. The Human Proteome Organisation (HUPO) has just been launched and at the time of the writing of this thesis, the frames of the project are being outlined. Mapping the entire proteomes corresponding to the model genomes of HUGO, means that enormous resources have to be available. Combining the efforts of many scientists around the world towards a common goal will make the process efficient and feasible.
**ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-DE</td>
<td>two-dimensional gel electrophoresis</td>
</tr>
<tr>
<td>2HS</td>
<td>two-hybrid system</td>
</tr>
<tr>
<td>ABD</td>
<td>albumin binding domain</td>
</tr>
<tr>
<td>ABP</td>
<td>albumin binding protein</td>
</tr>
<tr>
<td>APC/C</td>
<td>anaphase-promoting complex or cyclosome</td>
</tr>
<tr>
<td>BHK</td>
<td>baby hamster kidney cells</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>CHO</td>
<td>Chinese hamster ovary cells</td>
</tr>
<tr>
<td>COS</td>
<td>African green monkey kidney cells</td>
</tr>
<tr>
<td>DHFR</td>
<td>dihydrofolate reductase</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tag</td>
</tr>
<tr>
<td>GFP</td>
<td>green fluorescent protein</td>
</tr>
<tr>
<td>GST</td>
<td>glutathione S-transferase</td>
</tr>
<tr>
<td>HUGO</td>
<td>Human Genome Organisation</td>
</tr>
<tr>
<td>HUPO</td>
<td>Human Proteome Organisation</td>
</tr>
<tr>
<td>HSA</td>
<td>human serum albumin</td>
</tr>
<tr>
<td>IF</td>
<td>immunofluorescence</td>
</tr>
<tr>
<td>IgG</td>
<td>immunoglobulin G</td>
</tr>
<tr>
<td>IMAC</td>
<td>immobilized metal-ion affinity chromatography</td>
</tr>
<tr>
<td>MALDI-TOF</td>
<td>matrix-assisted laser desorption ionization – time of flight</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>Mb</td>
<td>megabase</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectroscopy</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>RACE</td>
<td>rapid amplification of cDNA ends.</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>RBS</td>
<td>ribosome-binding site</td>
</tr>
<tr>
<td>SAP</td>
<td>selection and amplification of phage</td>
</tr>
<tr>
<td>ScFv</td>
<td>single-chain variable region fragment</td>
</tr>
<tr>
<td>SCP1</td>
<td>synaptonemal complex protein 1</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>sodium dodecyl sulfate – polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>SELEX</td>
<td>systematic evolution of ligands by exponential enrichment</td>
</tr>
<tr>
<td>SIP</td>
<td>selectively infective phage</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>SPA</td>
<td>Staphylococcus aureus protein A</td>
</tr>
<tr>
<td>SPG</td>
<td>streptococcal protein G</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

First of all I would like to thank the entire group at the Dept of Biotechnology, and especially DNA corner™, for a friendly and stimulating atmosphere. For nice trips together to Paris, Sandhamn, Gotland, Väddö and Berlin and for nice parties every Christmas and at other times. There are some people that I would like to thank in particular:

Prof. Stefan Ståhl for being an ever-enthusiastic and optimistic supervisor. For giving expert guidance and, not the least, for being a good friend.

Prof. Mathias Uhlén who first got me interested in the field of bioscience by his inspiring lectures for the M. Sc. students. For giving me the chance to become a Ph D student at the DNA corner™ and for visionary leadership of the group.

My collaboration partners in Christer Höög’s group at the Dept. of Cell and Molecular Biology at the Karolinska institute, especially, Christer, Eva and Pia-Marie.

Dr. Magnus Larsson, my “x-jobb” supervisor who taught me almost everything I know about lab techniques. Also for being a very good partner at the lab bench while working four-handed with high-throughput projects.

Ronny Falk, my “x-jobb” student for the great efforts he has put into our common projects and for his never-ending optimism.

Sophia, P-Å and Jocke L for the protein club and for fruitful discussions on scientific matters and other things.

Jenny and Malin for being good friends, Maria E for interesting discussions on feminism and other important stuff, Stina for being a good partner teaching students and Anna B, my tap dancing colleague.

Olof and Erik for starting the music evenings and Tove, Per, Henrik, Amelie, Björn, Magnus, Anders, Nina, Martin, Susse, Maria W, Anna G and others for participating.

The Lucia choir for the fantastic opportunity to practise singing at the lab, Maria S, Lotta, Eric, Eva H, Fredrik, My, Kicki, Malin G, Linda, Eva B, Ulla and many others.

Old friends at the lab who has helped me a lot during the years: Marianne, Jocke Nilsson, Bertil, Elin, Christin, Sissela, Pelle, Anders Holmberg and others.

My friends, Daniel, Karin, Ullis, Lotta, Joel, Martin, Andreas, the members of EightCount, DAM, I heruedhil, Interaktiva uppsätningar, GOM, the former board of Fëa Livia and the Lindyhoppers and tap dancers in SSS. Thanks for the inspiration to other things than science.

Sara Ljungberg, for teaching me how to use my singing voice, for enormous encouragement and for being a very good friend.

My family: my brother Torbjörn who has also been a colleague and a “förbild”, my parents Astrid and Christian for love and making me believe in myself, Farmor for great support.

And most of all, my beloved Anders, who I share my life with. For everything.

This thesis has been financially supported by Teknikvetenskapliga forskningsrådet (TFR) and Stiftelsen för strategisk forskning (SSF). I would also like to thank Gålöstiftelsen (Sixten Gemzéus fond) and Kungliga vetenskapsakademien (Öfverdirektör Elis Sidenbladhs fond) for financial support.
REFERENCES


Functional proteomics: Generation and analysis of cDNA-encoded proteins


system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc. Natl. Acad. Sci. USA 97, 1143-1147.


