1. Introduction

Functional analysis of the proteins expressed in a specific tissue and/or stage of development is an essential step towards an understanding of biological processes. Initial attempts at sequencing the large and complex human genome were intentionally focused on expressed regions, as represented by cDNA repertoires (1,2). Meanwhile, expressed sequence tags (ESTs) for most human genes have been deposited in the nucleotide databases (3). However, only a minority of the proteins encoded by these sequences have yet been assigned a function (4).

For a full understanding of the function of a gene, information on the encoded protein's expression levels in different cell types, interactions, biochemical activities, structure, modifications, and localisation in the cell is required. The function of genes and the encoded proteins are studied by genetic and biochemical approaches.

By identifying mutations leading to genetic traits, conclusions can be drawn on the function of the affected genes. In a reverse approach, genes are deliberately deleted ('knocked out') or changed in their expression level, and the resulting phenotypic effect is studied (5,6).

1.1 Expression patterns

The function of genes can be inferred by their level of expression in tissues of different developmental stages, differentiation and disease status. Technology is available to correlate the functional status of a tissue with the expression of certain sets of genes. To determine the expression strength of a gene, the amount of transcript or the amount of protein in a cell are monitored.

On the transcriptional level, gene expression patterns are compared by DNA hybridisation or by sequencing approaches. The SAGE approach involves sequencing of bacterial clones containing concatenated 13 bp-tags of many different cDNAs (7). Tags identified in the sequences are matched to sequence database entries. When a sufficiently high number of clones is sequenced, the number of times a certain database entry is matched can be correlated to the expression strength of the respective gene. With this technique, the expression strength of a large number of genes in normal and cancer cells was determined and compared (8).

Expression patterns are being examined by DNA hybridisation. Thousands of cDNAs or oligonucleotides representing individual genes are arrayed on filter or glass slide supports. This was enabled by the development of devices that can array biological samples at high density with a high precision (9). To examine variation in gene expression, complex probes, generated by reverse transcription of RNA from different tissues and cell-lines, are hybridised on cDNA or oligonucleotide arrays. Differentially expressed genes show discriminative signal strength with complex probes from distinct tissues or cell lines.

With this technique, expression patterns of both resting and activated T-cells were compared using radioactively labelled probes which were hybridised on high-density cDNA filters (10). Expression differences in a tumour cell line before and after reversal of tumorigenicity were measured by hybridisation of fluorescently labelled probes in small volumes on cDNA micro-arrays printed on glass slides (11). High-density arrays of oligonucleotides representing yeast open reading frames were used to compare gene expression in yeast growing in rich and minimal media (12).

Oligonucleotide fingerprinting (or OligoFingerprinting) is another DNA hybridisation approach applicable to monitor expression patterns (13-15). This technique involves hybridisation with short oligonucleotide probes to profile gene expression patterns of cDNA libraries. Sets of short oligonucleotides are subsequently hybridised on high-density cDNA arrays. Clones that hybridise with the same subsets of oligonucleotides are grouped into clusters. The size of a cluster correlates with the expression strength of the respective gene. By oligonucleotide fingerprinting cDNA libraries of different developmental or disease stages, or of different tissues, expression levels of thousands of genes are compared in parallel. The expression levels of genes in human infant brain were measured by this method (14), and gene expression in nine day and twelve day mouse embryos was compared (16).

1.2 Protein expression

Expression patterns can be analysed on the translational level. Two-dimensional electrophoresis (2D-PAGE) is used to map protein extracts at high resolution (17) and sequence information on protein spots in 2D-gels is obtained by mass spectrometry (18). Protein extracts are obtained by subcellular fractionation of tissues or cell lines. Recombinant DNA technology has created an alternative route to generate proteins of interest in high yield and purity. The rapid progress in characterising genes and mRNAs (expressed sequence tags, ESTs) as a result of the Human Genome Project, creates a need for expression of the encoded proteins. Several options exist for overexpressing cloned genes in host cells or *in vitro*, and for detection and purification of the expression products.

1.2.1 Expression of fusion proteins

To facilitate the purification of the expression product, protein sequences are fused to protein or peptide tags. A large number of tags have been introduced for this purpose. Fusions with glutathione S-transferase or maltose binding protein enable affinity chromatography on immobilised glutathione or amylose, and furthermore can improve the solubility of the fusion partner (19,20). A number of binding epitopes of monoclonal antibodies have been introduced as tags, e.g. epitopes of antibodies against c-myc (21), influenza virus hemagglutinin (22) or vesicular stomatitis virus glycoprotein (23). These tags are around ten amino acids long, and allow highly specific detection of the expression product in cellular extracts. On the other hand, affinity purification with immobilised antibodies usually involves elution at a low or high pH, which can denature the target protein. An exception is the epitope of a monoclonal antibody directed against protein C, a vitamin K-dependent plasma zymogen (24). Binding of the protein C antibody is Ca^{2+}-dependent, and protein can be mildly eluted with Ca^{2+}-chelating reagents.

The StrepII-tag is a nine amino acid peptide with affinity for streptavidin, that allows purification on a matrix conjugated with modified streptavidin (25,26). Target proteins are eluted with biotin.

Metal chelate affinity chromatography for protein purification is based on the ability of amino acids such as histidine, acting as electron donors, to bind reversibly to transition-metal ions that have been immobilised by a chelating group covalently bound to a solid support (27). Recombinant proteins engineered to have six consecutive histidine residues on either the amino or carboxyl-terminus can be purified using a resin containing Ni^{2+} ions that have been immobilised by covalently attached nitrilotriacetic acid (Ni-NTA, ref. 28). This technique can be performed with either native or denatured protein. Ni-NTA conjugated to marker enzymes and antibodies directed against the His_6-tag is available and can be used to detect fusion proteins in cellular extracts.

1.2.2 Expression systems

When choosing an expression system, the following parameters have to be considered: the amount of recombinant protein that can be expressed, the time needed to set up expression and the ability of correct folding and posttranslational modification. The most widely applied expression systems are those using *Escherichia coli* as host cell (29-31). The genetics and biochemistry of this organism are probably the best understood of any known organism. *E. coli* can be easily genetically manipulated, has a short doubling time and can be cultivated on inexpensive media. Strategies for gene expression in *E. coli* include intracellular expression and secretion of the protein into the periplasmic space. Expression in *E. coli* has the disadvantage that eukaryotic proteins are often not correctly folded, leading to the formation of dense, insoluble aggregates called inclusion bodies. Furthermore they lack post-translational modifications as phosphorylation or glycosylation. Inclusion bodies are easily separated from soluble cell components by centrifugation and can be solubilised by chaotropic salts or detergents. His_6-tagged recombinant proteins can be purified under denaturing conditions by metal chelate chromatography and used for immunisation of animals. However, for functional analysis the expressed proteins have to be in their native state.

Yeast is another unicellular organism that retains advantages of bacterial systems as ease of manipulation and growth. Yeast possesses a eukaryotic secretory pathway and can effect disulphide bond formation, proteolytic maturation, N- and O-linked glycosylation and other post-translational modifications of many mammalian proteins. For *Saccharomyces cerevisiae*, a wealth of information on genetics, molecular biology and physiology has been accumulated (32,33), and the whole genome has been sequenced (34). A variety of promoters, plasmids and selectable markers is available (35). On the other hand, secreted proteins are often hyperglycosylated and, except for a few examples, yields of heterologous protein only reach a maximum of 1–5% of total protein, even with a strong promoter. This limitation is now relieved by an increasing number of alternative non-*Saccharomyces* yeast which exhibit favourable properties as protein production organisms, e.g., the methylotroph *Pichia pastoris* (36).

Protein expression in cultured animal cells is more demanding than in micro-organisms in terms of growth conditions and establishment of expression. Insect cells transfected by baculovirus vectors have become popular expression hosts, because cells are easily cultured and can be grown in bioreactors, and high expression levels of up to 25% of total cellular protein can be obtained (37-39). Proteins are expressed intracellularly or secreted into the medium and are usually folded correctly. Because of the size of the baculovirus genome, recombinant baculovirus is generated *via* homologous recombination with a plasmid containing the gene of interest. The efficiency of this step has been improved (40-42) but the generation of recombinant baculovirus is still long winded in comparison to *E. coli* or yeast systems. Direct cloning into baculovirus vectors has been introduced by inserting a rare cutter restriction site into the baculovirus genome (43,44). Both insect cells and yeast are capable of correctly folding eukaryotic proteins and formation of disulphide bonds. They can effect post-translational modifications, but these often are not identical to those found in proteins expressed in mammalian cells.

Mammalian proteins expressed in mammalian cell culture can be assumed to be correctly folded and processed. For protein expression, cell lines are transfected stably or transiently. Transient transfection of a transformed green monkey kidney cell line (COS-7 cells, ref. 45), that expresses the SV40 large-T antigen is well established (46). Upon transfection of a plasmid carrying the SV40 origin, the plasmid becomes amplified, and strong protein expression continues for about a week. Microgram amounts of protein can be achieved. Transient expression is difficult to scale up, and for expression in bioreactors, stably transfected cell lines have to be established (47). Stably transfected cell lines are obtained by selecting for integration of plasmid DNA into the host chromosome. After transfection of the target gene linked to a selectable marker gene, high-level expression of the target gene is obtained by selecting for amplification of the marker gene. This may take several month, after which milligram amounts of protein can be expressed in bioreactors.

In vitro transcription-translation is an effective alternative to expression in host cells. A PCR product or a plasmid preparation of a cloned cDNA sequence is used as template for mRNA synthesis by T7 or SP6 phage RNA polymerases (48), followed by translation in cellular extracts. *In vitro* transcription-translation in wheat germ extracts (49) or rabbit reticulocyte lysates (50) has been used to express a multitude of functional proteins. The expressed protein is labelled by incorporation of radioactive or biotinylated amino acids (51). Extracts from *E. coli* have been used for *in vitro* translation for a long time (52), and have been turned into highly efficient systems by introducing membrane-fitted bioreactors for exchange of nucleotides and amino acids and removal of expression product (53,54). In comparison to expression in *E. coli* cells, *in vitro* expression product can be purified more easily, since no cell harvesting and cell lysis steps are involved, and less contaminating protein is present.

Furthermore, proteins toxic to *E. coli* cells may be expressible *in vitro*, and modified amino acids can be introduced into the expression product.

1.3 Expression libraries to study protein interactions

Recombinant expression of a protein of interest can be a prerequisite for the analysis of interactions with other proteins. For studying protein-protein interactions, a variety of techniques is available. The yeast two-hybrid system selects for intracellular interaction of a protein with the expression product of a cloned cDNA fragment (55,56). In coprecipitation or 'pulldown' experiments, an immobilised recombinant protein is used to capture interacting proteins in cell extracts. Expression cloning or interaction cloning involves screening of cDNA libraries cloned in bacterial expression vectors, to identify clones that express proteins interacting with the target protein (57,58). Expression libraries cloned in λ phage expression vectors (59) were developed to identify clones by screening with antibodies (60) and were also screened for proteins binding to nucleic acids (57,58), other proteins (61,62) or metabolites (63), and for expression products that are recognised by modifying enzymes (64). As an alternative to phage vectors, plasmid vectors were used as they offer increased expression levels (65,66). Random cDNA clones often contain out-of-frame fusions to the vector encoded translated sequence. To circumvent this problem, runs of adenines or thymines were introduced before the cDNA insert to effect slippage of the *E. coli* RNA polymerase, and thereby correct out-of-frame fusions (67).

Expression cloning with mammalian cDNA expression libraries involving transient expression was done in COS-7 cells. This technique has been widely used for the cloning of receptors or cell surface antigens. Transfected cells that express a cDNA of interest were selected on antibody coated dishes (68,69). Fluorescence-activated cell sorters (FACS) were applied to collect cells that express a cDNA of interest (70). Cytokine receptor cDNAs have been isolated by screening for cytokine binding (71-73).

1.4 Arrayed DNA libraries and high-density grids

The use of arrayed DNA libraries has become widely established in genome analysis laboratories (13,74,75). Arraying of DNA libraries was enabled by the development of automated systems that are able to pick large numbers of clones from spreads on agar plates into microtitre plates (76). For the efficient characterisation of the arrayed clones by DNA hybridisation, high-density gridding of clones or PCR products on DNA-binding membranes was developed (75). Experimental data generated on high-density grids of arrayed libraries are directly related to clones permanently stored frozen in microtitre plates. This enables high throughput characterisation of large DNA libraries by DNA hybridisation. YAC, cosmid and P1 clone contigs spanning the whole *Schizosaccharomyces pombe* genome were generated by this technology (15,77). Another important advantage of arrayed libraries is that they can be replicated without limitation, and copies can be distributed to other laboratories. Thereby the arrayed clones can be characterised by experimental data generated by different laboratories. Data is generated by various experimental strategies including DNA sequencing, DNA hybridisation and *in situ* hybridisation. Hybridisation probes may be obtained from single clones, or as complex mixtures representing tissues or cell lines of interest. They include oligonucleotides, cDNAs or genomic sequences. DNA probes can identify clones representing a specific gene, a gene family or genes expressed in a certain tissue, cell line or disease state, or clones located on a certain genomic region. Oligonucleotide fingerprinting involves hybridisation with short oligonucleotide probes to profile gene expression patterns of cDNA libraries (see 1.1). Arrayed library clones may be used as DNA probes to screen DNA libraries, or for *in situ* hybridisation to detect expression patterns in developmental, diseased and normal tissues. To supply different laboratories with arrayed clone libraries, and to collect and integrate the data that is generated, organisations as the IMAGE consortium (78) or the Resource Centre of the German Human Genome Project (RZPD, http://www.rzpd.de, ref. 79) have been established.

1.5 Systematic protein expression

The arrayed DNA library approach integrates data on arrayed library clones generated by DNA sequencing and hybridisation techniques. By the development of expression libraries, techniques for DNA library screening have been extended from DNA-based to protein-based techniques as antibody screening. The establishment of arrayed expression clone libraries enables the integration of data generated by DNA-based techniques and protein-based techniques. Systematic protein expression from arrayed cDNA library clones has been described for *in vitro* and mammalian cell expression systems. A panel of cDNA clones were transfected into mammalian cells, and overexpressed proteins were subjected to 2D-PAGE

(80). This allowed matching of cDNA clones to 2D-PAGE patterns from human cells. Another method for characterising the expression products of cDNA clones involved pooling of clones followed by *in vitro* transcription and translation in reticulocyte lysate. The radioactively labelled expression products were separated by 2D-PAGE. By using overlapping pools, spots on 2D-gels were assigned to individual cDNA clones (81-83).

These approaches, albeit characterising novel gene products by gel electrophoresis, cannot provide larger amounts of purified protein necessary for biochemical and biophysical studies. As aforementioned, larger amounts of protein can be obtained by expression in *E. coli*, yeast (e.g. *Pichia pastoris*), expression in insect cells with the baculovirus system or stably transfected mammalian cell lines. *In vitro* expression in prokaryotic extracts could become a cost-effective alternative, used in batch reactions for large numbers of transcripts, and with the option to scale up production in protein bioreactors. While expression in insect or mammalian cells may be too time-consuming to set up for large numbers of clones, expression in *E. coli* and yeast remains under scrutiny. These systems may be well suited to generate expression products for large numbers of clones in relatively high amounts.

In the approach described here, the aim was to construct arrayed cDNA expression libraries that are suitable for standard DNA analysis but, in addition, could serve as a source of recombinant proteins and could be characterised by protein detection screening. For library generation, cDNA molecules are directly cloned into a vector for the expression of fusion proteins. This has some inherent consequences. Protein coding sequences randomly fused to an expression vector are expected to be out of frame in two thirds of cases. cDNA generated by oligo(dT) priming generally comprises the original stop codon, but may, subject to transcript length, lack the start. On the other hand, if a full-length cDNA is obtained, no N-terminal fusion protein will be expressed if the 5'-untranslated region contains a stop codon in the frame of the encoded protein.

Using arrayed expression libraries offers the possibility to screen the whole library at once for those clones that express recombinant protein. The selective detection of *E. coli* clones expressing His$_6$-tag fusion proteins in the correct reading frame by an antibody directed against this tag has been described (84). This technique relies on the degradation of misfolded and short proteins and peptides in the *E. coli* cytoplasm. Abnormal or misfolded proteins are cleaved by energy-dependent proteases, and short peptides of 6–15 amino acids are released. These peptides are subsequently degraded by energy-independent proteases and peptidases (85). Translation of cDNA inserts in an incorrect reading frame usually leads to the expression of short, and therefore unstable peptides because of the high abundance of stop codons in the alternative reading frames.

1.6 Protein characterisation by mass spectrometry

Recombinant protein expression is routinely monitored by SDS-PAGE. By comparison with molecular weight standards, the molecular mass of the expression product can be estimated. Because of the low accuracy of the determined molecular mass, it is not possible to detect single amino acid changes or the absence of small parts of the protein resulting from proteolytic cleavage or from incomplete transcription or translation.

The development of matrix assisted laser desorption/ionisation (MALDI) mass spectrometry has led to significant advances to the characterisation of biopolymers such as proteins or peptide mixtures (86). MALDI mass spectrometry is being used for the identification of unknown proteins and also for the quality control of recombinant expression products. Proteases or chemicals are utilised that cleave proteins at predictable sites, e.g., trypsin, lysyl endopeptidase or cyanogen bromide. The masses of the released peptides are determined by mass spectrometry and compared to those predicted from the sequence of the protein, if known. For the identification of unknown proteins, databases have been developed that contain the predicted peptide masses of all available protein sequences cleaved with a variety of proteases and cyanogen bromide (87). The unknown protein is cleaved and the masses of the generated peptides are determined. By matching these masses with those predicted from protein database sequences, the unknown protein can be identified.

1.7 Antibodies

Antibodies are indispensable tools for the analysis of proteins. Immunoglobulin G (IgG) is the main isotype of mammalian antibodies. IgG molecules consist of two identical pairs of polypeptide chains, the heavy and the light chain, connected by disulphide bonds. Heavy and light chain comprise constant and variable regions. The variable regions encode loops of highly diverse structure at the tip of the antibody. The diversity of antibody populations comprises binding specificities for millions of binding epitopes.

Antibodies have led to the development of various techniques for protein analysis. These include immunoblotting, ELISA, immunoprecipitation and affinity purification, immunohistochemistry and flow cytometry.

Immunoblotting is used to detect specific antigens in protein mixtures separated by electrophoresis. ELISA is a sensitive method to detect and quantify antigens in crude lysates or other solutions. In immunoprecipitation, antigens are isolated from protein mixtures using antibodies. Antibodies are immobilised to a matrix support to capture the protein of interest. Alternatively, complexes of a protein of interest with a specific primary antibody are formed, which are subsequently immunoprecipitated with secondary antibodies directed against the constant region of the primary antibody.

Immunohistochemistry reveals the expression pattern of proteins in tissue specimens. Fluorescence microscopy is used to locate antigens within cells. Flow cytometry involves the surface staining of cells with fluorescently labelled antibodies. In a fluorescence-activated cell sorter (FACS), individual cells flow past a fluorescence excitation and detection device, and are sorted according to the measured fluorescence. A typical application of flow cytometry is the immunofluorescence staining of an extracellular antigen to count or select cells expressing this antigen in cell populations, or to quantify its expression.

Antibodies can be generated by a variety of techniques. Antigens injected into the body of animals trigger an immune response. Immunoglobulins with binding specificity for the antigen are released into the blood serum. Adjuvant can be injected together with the antigen to enhance the immune response (88). The serum of immunised animals contains a polyclonal mixture of antibodies directed against different epitopes of the antigen.

Monoclonal antibodies were made available by fusion of immune B cells from the spleen of immunised animals with tumour cells to produce hybridomas, followed by selection of clones that express a desired antibody (89). Monoclonal antibodies are advantageous in many applications, because they can be reproducibly produced in large amounts and are highly specific for a single epitope. Monoclonal antibodies that bind to functional domains can be used to alter protein activities, e.g. monoclonal antibodies that represent specific antagonists or agonists of ligand-receptor interactions. On the other hand, a large effort is necessary for the establishment of monoclonal antibodies. Polyclonal antibodies are well suited for many applications as immunoblots or ELISA, but background problems can arise since they contain diverse binding specificities. Furthermore, only a limited amount of antiserum can be obtained from a single immunised animal, and immunisation of different animals yields antisera of varying quality.

As an alternative to the immunisation of animals, specific antibody fragments displayed on bacteriophages can be selected from large repertoires (90,91). Highly diverse repertoires of antibodies are generated by cloning the variable regions of naturally occurring immunoglobulins, or by substituting variable regions in immunoglobulin sequences with synthetic randomised sequences. Libraries of heavy and light chain genes can be generated independently in *E. coli* vectors. By recombination of light chain and heavy chain libraries, highly diverse repertoires are generated. Antibodies are expressed in *E. coli* cells as fusions with a filamentous phage coat protein and phage presenting antibody fragments fused to the coat protein are released. Phage presenting antibody fragments against an antigen of interest are selected by several rounds of binding and elution on immobilised antigen. Antibodies cloned by this method may be expressed in large quantities in *E. coli* or other hosts.

2. Objective

Up to now, no technique has been available to go directly from DNA sequence information on individual clones to protein products and back again at a whole genome level. The objective of this study is the establishment of arrayed expression libraries for the integration of DNA-based and protein-based experimental data. Libraries should be amenable to DNA-based techniques as DNA sequencing, oligonucleotide fingerprinting and other DNA hybridisation strategies, in addition to protein-based techniques as antibody screening. Clonal expression products should be directly utilised as antigens for the generation of antibodies, and also for their biological characterisation. Therefore, proteins should be expressed as fusions to peptide or protein tags for their efficient detection and purification by affinity chromatography. The choice of the expression system is a trade-off between hosts with different properties, as being easy to grow and to alter genetically, offering high protein yield or effecting correct folding and post-translational modifications.

In the approach described here, proteins are clonally expressed from arrayed cDNA libraries. This makes translated gene products directly amenable to high throughput experimentation and generates a direct link between protein, expression and sequence data. A bacterial expression system was chosen that allows expression of large numbers of proteins at relatively low cost.

The first aim of this work was to establish arrayed cDNA expression libraries that could be characterised by DNA and protein-based techniques as DNA hybridisation and antibody screening in parallel. Methods for differentiation between clones expressing recombinant protein and non-expression clones should be established, leading to arrayed libraries of expression clones. In addition to screening of the library on high-density grids, methods for high-throughput protein expression in liquid culture and characterisation in terms of size, expression strength and solubility should be established. These data will be integrated with results from DNA and antibody screening, and finally oligonucleotide fingerprinting, to establish a catalogue of identified and characterised expression clones.