6. Discussion

This study describes arrayed cDNA libraries as a source of clonally expressed recombinant proteins which can be directly linked to clones characterised and identified by DNA hybridisation or sequencing. cDNA expression libraries were screened for clonal protein expression on automatically gridded high-density filters with an antibody directed against the vector-encoded tag sequence of the expressed proteins. Putative expression clones detected in the hEx1 library were rearrayed into a sub-library. For the characterisation of expression products of large numbers of clones, the expression and purification of proteins in microtitre plates was established. The hEx1 library was screened with cDNA probes of human genes. Among the positives, putative expression clones detected by antibody screening were selected, and protein expression was verified. Identified expression clones were directly used to express and purify recombinant human proteins. Antibodies against human GAPDH and HSP90α were used to screen the hEx1 library in parallel with DNA probes of these genes. These experiments demonstrate that DNA-based and protein-based experimental data can be integrated by using arrayed cDNA expression libraries, creating a direct link between the gene catalogue and a gene product catalogue. A gene product catalogue containing recombinant protein of large numbers of genes could facilitate the characterisation of large numbers of proteins, and furthermore allow the screening of protein libraries for biological activities or binding characteristics of interest.

6.1 Arrayed cDNA expression libraries

6.1.1 Robot technology and arrayed libraries

The application of high-density arrayed clone filters has now been extended to antibody screening of expression libraries, and potentially to screening for biological functions of the expression products. The generation of an arrayed expression library may be regarded as an advance over currently available cDNA expression libraries, which are obtained as mixtures of clones (60). For isolation of clones from these libraries, usually multiple rounds of screening, picking and rescreening are necessary. In contrast, positive signals on filters of gridded libraries are directly related to stored clones and information generated by DNA-based

techniques. Based on this principle, the hEx1 library was subsequently screened with various antibodies and DNA probes. Expression clones were detected by both DNA probes and antibodies.

The expression system and the arraying technology used here offer a direct link from clones to purified protein. Clones identified by DNA hybridisation or sequencing among the putative expression clones in the hEx1 or mKd1 libraries can be directly used to express and purify the encoded protein. This saves the investigator subcloning identified cDNA sequences into an expression vector, which can become a rate limiting step if large numbers of proteins are to be analysed.

6.1.2 cDNA library construction

Oligo(dT) and random primed libraries

The mKd1 and hEx1 libraries were generated by oligo(dT) priming according to Gubler and Hoffman (101). Therefore, cDNA inserts are biased to 3'-ends of genes, and N-terminal parts of larger proteins may be missing. Random primers could be used in the future instead of oligo(dT) to synthesise cDNA that contains 3' and 5'-ends of genes with similar probability. Random primed libraries should be used in addition to oligo(dT) primed libraries to include a maximum number of epitopes for antibody screening. A human fetal brain library has been constructed in the pQE30NST vector using random primers with tails containing restriction sites for directed cloning (B. Korn, unpublished results). This libarary will be arrayed and distributed by RZPD in due course to complement the limited representation of N-terminal parts of large proteins in the hEx1 library.

Insert size

The average insert size of the mouse mKd1 and human hEx1 libraries was determined as 1.7 kbp and 1.5 kbp, respectively. For comparison, 6,642 human mRNA sequences labelled as 'complete codes' were retrieved from GenBank, and the average length was determined as 2.4 kbp. This is a rough estimate of the average transcript size in human cells. The actual value might be larger, because complete sequences of short transcripts are more easily obtained than of long transcripts, and therefore complete sequences of short transcripts may be over-represented in the GenBank database. Incomplete cDNA molecules are obtained if mRNA is degraded during preparation, or if mRNA molecules contain regions which are not readily

reverse transcribed. Most cDNA clones of long transcripts are therefore expected to be incomplete. This expectation was confirmed during the analysis of GAPDH and HSP90α clones and of 96 random putative expression clones in the hEx1 library. The GAPDH transcript is short (1.2 kbp), and correspondingly many clones were found to contain the full coding region. For HSP90α, which has a 2.9 kbp transcript, only 2 out of 35 clones contained the full coding region (Figure 14). The characterisation of 96 clones from the hEx1 library revealed 58 clones corresponding to protein sequences in the SWISS-PROT and TrEMBL databases. 38 (66%) of these clones included the start of a protein database sequence, and were therefore assumed to contain the full protein coding region. cDNA sequences of full-length clones matched protein database sequences of 35 kd average size – in contrast to clones containing truncated coding regions, which matched protein sequences of 61 kd average size.

Library size

The human genome has been estimated to contain 50,000 to 100,000 genes (117). A typical mammalian cell expresses approximately 10,000 genes. Estimates for the number of genes expressed in brain tissue, which is composed of a large number of different cell types, are similar to estimates for the total number of genes (117,118).

The abundance of transcripts of different genes is highly diverse, and range from several percent of the mRNA population for genes as α-tubulin or elongation α1-factor in human infant brain (1) to a few molecules per cell. The number of genes represented in a cDNA library is smaller than the number of clones, as many genes will be represented by more than one clone. The probability that a gene is found in a cDNA library depends on the number of clones in the library and on the abundance of the transcript in the mRNA population. In a non-normalised library, the abundance of a transcript in an mRNA population is generally related to the frequency of cDNA clones to this transcript in a library. If a transcript is rare, e.g. represented only once in 10^5 mRNA molecules, the chances of finding the corresponding cDNA in a library of 200,000 clones (approximately the size of the hEx1 library) is $1-\left(1-10^{-5}\right)^{200,000} = 86\%$. The hEx1 library is therefore considered to contain a major part of transcripts expressed in human fetal brain with an abundance of at least 10^{-5}. However only a fraction of the library consists of expression clones. The rearrayed sub-library of hEx1 contains 37,830 putative expression clones. Consequently, expression clones corresponding to transcripts of 10^{-5} abundance are found with a probability of only 31%.

To increase the number of genes represented in the rearrayed hEx1 sub-library, the number of clones might be further increased. Alternatively, cDNA library normalisation could be used to decrease the average number of clones per gene (115,119). To enable the expression of cDNA inserts that are fused out of frame to the vector-encoded start codon, a vectors have been used that contain runs of adenines or thymines before the insert, leading to slippage during transcription which corrects the frame-shift in a subset of transcripts (67).

6.1.3 Detection of expression clones

The selective detection of expression clones by the RGS·His antibody is apparently caused by the instability of short, unfolded polypeptides in *E. coli* cells. When a cDNA sequence is expressed in an incorrect reading frame, the resulting polypeptide will generally be short because of the high frequency of stop codons in non-coding reading frames, and is therefore not expected to fold into a stable structure. Such polypeptides are likely to be degraded within the host cell leaving the clone undetectable by the RGS·His antibody (85). This was confirmed by the analysis of GAPDH and HSP90α clones. 90% of clones with inserts in an incorrect reading frame were not detected by the RGS·His antibody, while clones expressing His_6-tagged fusion proteins were reliably detected.

The mKd1 and hEx1 cDNA libraries were prepared in a bacterial expression vector. At most, one third of clones were expected to contain inserts expressed in the correct reading frame as His_6-tag fusion proteins. The number of expression clones is further reduced because of cDNA inserts that comprise 5'-untranslated sequences with stop codons before the open reading frame or only 3'-untranslated sequences, or encode proteins that are toxic or rapidly degraded in *E. coli* cells. Screening of high-density protein filters of the mKd1 and hEx1 libraries with the RGS·His antibody detected about 20% of clones, with different signal intensities, grouped into level one (weak) to three (strong). Analysis of twelve mKd1 clones detected with intensity level one, two or three showed a high proportion of clones expressing recombinant protein of at least 15 kd size, which were therefore assumed to contain inserts in the correct reading frame. This proportion was higher (83%) for intensity three than for level one (42%). It was concluded that the probability of clones expressing recombinant protein is correlated to the intensity of this RGS·His detection signal.

6.2 Rearraying of the hEx1 library

6.2.1 Expression products of 96 random clones

Screening with the RGS·His antibody, in combination with robot technology led to the generation of an hEx1 sub-library containing 37,830 putative expression clones.
96 randomly selected clones, which were detected with medium or high signal intensity, were subjected to detailed analysis of their protein products and DNA sequences. For 63 clones (66%), protein expression of at least 15 kd size was visible in SDS-PAGE of whole cellular extracts. Metal chelate affinity purification revealed expression in six additional clones. The solubility of the expression products was tested. 15 (24%) of 63 expression products were found in the soluble fraction. The remaining expression products were assumed to form inclusion bodies. Using these proteins for biological assays would require to establish their denaturation and subsequent refolding. For refolding of proteins a suitable protocol has to be determined empirically for each protein (120). His_6-tag fusion proteins can be purified under denaturing conditions and used for the generation of antibodies by immunisation of animals, or by selection from antibody phage display repertoires (see 1.7).

6.2.2 DNA sequence analysis

DNA sequences of the 5'-ends of the cDNA inserts were generated and compared to the combined SWISS-PROT and TrEMBL databases (113). The sequences of 58 clones were found to match human protein sequences in these databases. Among these clones a percentage of 66% comprised a complete protein coding region. As expected, full-length sequences matched smaller protein sequences in the databases (average 35 kd) than sequences lacking the N-terminus (average 61 kd).
38 (66%) of the 58 known protein coding sequences were fused in frame to the vector-encoded His_6-tag. Almost all proteins generated by translation of incorrect reading frames were shorter than 25 kd, while most inserts in the correct reading frame yielded proteins larger than 25 kd (Figure 5). Six clones with inserts in an incorrect reading frame expressed proteins of more than 14 kd size (estimated from SDS-PAGE). This number may be extrapolated to approximately 2% of clones in the hEx1 library before rearraying, which is a larger percentage than would be expected by statistical calculus. If a random distribution of amino acids was assumed for non-coding reading frames, stop codons would occur at a frequency of three in

64. Consequently, the probability of a random open reading frame to comprise more than 130 codons, corresponding to a protein of more than 14 kd, would be $(61/64)^{130} = 0.19\%$. A possible explanation for the discrepancy between the calculated and observed values is that a statistical distribution of stop codons cannot be generally assumed, as regions of reduced complexity, e.g. repetitive regions, may contain lower (or higher) frequencies of stop codons.

6.2.3 Protein sizes predicted from DNA sequences and estimated from SDS-PAGE

35 (60%) of the 58 clones with sequences matching SWISS-PROT or TrEMBL database entries expressed a recombinant protein in the correct reading frame that was visible in SDS-PAGE of whole cellular proteins and/or in SDS-PAGE of metal affinity purified protein. The sizes of the expressed proteins were estimated by SDS-PAGE. For 27 of 35 expression products, the estimated sizes matched the predicted sizes with at least the reported accuracy for this method of ±10% (121), while for eight clones, the deviation of the observed protein size from the predicted size was larger. One clone clearly expressed a truncated protein. The sizes of four expression products were predicted less than 20 kd, where the resolution of the gel system was limited which may have led to the observed deviations.

Availability of the hEx1 library

DNA and protein filters of the rearrayed hEx1 library for screening with DNA or antibody probes were prepared by the RZPD and are available on request. DNA hybridisation and antibody screening results were entered into the Primary Database of the RZPD. Sequence information and data on expression products will be entered in due course.

The rearrayed hEx1 library represents a resource of protein expression clones for a large portion of the human genes. Approximately 25% of recombinant proteins are expressed in soluble form, and are therefore directly amenable to biochemical and biophysical experimentation. Expression products that form inclusion bodies can be used for the generation of antibodies or refolded. Despite the limitations of the cDNA construction strategy used, two thirds of clones were found to contain a complete protein coding region.

6.3 Identification of specific expression clones

6.3.1 Screening for expression clones of nine human genes

The hEx1 library was screened by DNA hybridisation with probes for nine different genes. By correlation with the RGS·His antibody screening results, expression clones for seven genes were identified. For GAPDH and calmodulin, almost all clones identified contained the complete coding region, and expressed protein in soluble form. The protein products of a GAPDH and a calmodulin clone were shown to be biologically functional in enzyme assays. Clones of the larger HSP90α and HSP90β transcripts were mostly incomplete. Expression strength and solubility of the expression products decreased with increasing HSP90 insert size. It is possible that larger inserts contained sequences detrimental to expression in *E. coli*, which were missing in the shorter inserts.

A cell lysis buffer containing 1.5% of the detergent sarkosyl was used for the solubilisation of fusion proteins of HSP90β and subunit IV of cytochrome c oxidase (COX4), which were insoluble under non-denaturing conditions. HSP90β and COX4 were eluted with non-denaturing buffers from metal chelate affinity columns, suggesting that lysis in sarkosyl may solubilise aggregates of certain proteins without denaturation, which had been described for glutathione S-transferase fusion proteins before (109).

For the anion channel VDAC1, none of the positives in the hEx1 library were detected by the RGS·His antibody, possibly because this transmembrane protein cannot be expressed in *E. coli*. For RXRβ, four of 16 positives were also detected by the RGS·His antibody, but only one clone was found to contain an RXRβ insert in the correct reading frame and to weakly express a protein of the expected size. Toxicity of RXRβ fusion proteins to *E. coli* cells could explain the weak expression and the small number of expression clones detected by the RGS·His antibody. Correspondingly, no increase in the optical density in cultures of the RXRβ expression clone was observed upon induction of expression, suggesting that cells had stopped dividing.

6.3.2 GAPDH and HSP90α expression clones

The detection of expression clones with DNA probes and antibodies was analysed in detail for two example proteins, GAPDH and HSP90α. Screening for clones expressing these two proteins produced various categories of clones that allowed an evaluation of the technique.
Table 8 and Figure 11 summarise the yields of clones in the different categories, reflecting abundance of mRNA messages and efficiency of protein expression. In a non-normalised cDNA library like hEx1, the frequency of occurrence of gene-specific clones depends on gene expression levels. The two example genes, GAPDH and HSP90α, are both transcribed at relatively high levels. GAPDH as a typical house keeping gene, encoding an enzyme central to the cellular metabolism, was found to be represented by 237 (0.29%) clones, whereas for HSP90α, 56 (0.07%) clones were found, indicating four times less abundance of mRNA messages than for GAPDH.

Detection of expression clones by the RGS·His antibody and protein-specific antibodies

25% of the GAPDH and HSP90α clones were detected by the RGS·His antibody. Antibodies against GAPDH and HSP90α were independently used to detect expression clones. 61% of the GAPDH and 72% of the HSP90α clones detected by the RGS·His antibody were also detected with the protein-specific antibodies. Expression of GAPDH or HSP90α fusion proteins by clones detected by both the RGS·His antibody and a protein-specific antibody was verified by SDS-PAGE of whole cellular proteins and western blotting (Figure 12, Figure 13).
Clones detected by the RGS·His antibody but not by the specific antibody were mostly found to contain inserts in incorrect reading frames. Two clones expressing truncated, C-terminal parts of GAPDH were also found in this category. These expression products were only poorly detected in western blot by the GAPDH antibody. It is possible that the GAPDH antibody mainly recognises epitopes in the N-terminal half of GAPDH. Taking into account that clones expressing truncated His_6-tagged GAPDH fusion proteins are not detected by the GAPDH antibody, it was estimated that approximately 72% of both GAPDH and HSP90α clones detected by the RGS·His antibody express His_6-tagged fusion proteins in the correct reading frame. The remaining 28% of RGS·His positive clones may have expressed short polypeptides which were not completely degraded in the *E. coli* cells. Five GAPDH and two HSP90α clones detected by the RGS·His antibody on high-density protein filters containing inserts in an incorrect reading frame were analysed by western-blotting with the RGS·His

antibody (Figure 12 lanes 13–17, Figure 13 lanes 6,7). His_6-tag fusion proteins were not or only poorly detected on these blots. It is possible that the degradation of small His_6-tag fusion proteins was more efficient during protein expression in liquid culture than in the course of the preparation of high-density protein filters. The reliability of the RGS·His antibody for detecting expression clones was confirmed as only 9% of clones with GAPDH or HSP90α inserts cloned in an incorrect reading frame were recognised by this antibody. On the other hand, all GAPDH and HSP90α expression clones detected by the protein-specific antibodies were reliably detected as expressing His_6-tagged fusion protein.

Percentage of expression clones

25% of GAPDH and HSP90α clones were detected by the RGS·His antibody, and estimated 72% of these clones expressed GAPDH or HSP90α fusion proteins. This is less than the 33% that would, theoretically, be expected to contain inserts in the correct reading frame. The difference may be caused by small cDNA inserts containing none, or only a small part of an open reading frame. Moreover, incomplete repression in the non-induced state may have led to impaired growth of expression clones, and consequently, non-expression clones may have been preferably obtained when the cDNA library was prepared.

Second screening with the HSP90α antibody

Screening of the hEx1 library with the HSP90α antibody initially did not reveal all clones that expressed His_6-HSP90α fusion proteins. A second screening involved longer substrate development times for maximal colour intensity, and detected clones missed in the first screening which were also detected with the HSP90α DNA probe.

Expression of HSP90α from internal start sites

A considerable number of clones positive with the HSP90α DNA probe and the HSP90α antibody were not detected by the RGS·His antibody, indicating the expression of HSP90α sequences without a His_6-tag (Figure 13 E and
Table 8). On western blots, multiple bands representing differently sized expression products were detected for these clone (Figure 13 E). The sizes of the largest proteins expressed by the different clones on the anti-HSP90α western blot correlate with the largest possible translation products of their cDNA inserts (Figure 14). Larger inserts would contain the start sites present

in smaller inserts plus additional start sites. Correspondingly, proteins expressed by clones with smaller inserts were also expressed by clones with larger insert.

In a second screening with the HSP90α antibody, even more HSP90α clones negative with the RGS·His antibody were detected. The sequenced clones were shown to have inserts in an incorrect reading frame including the C-terminal part of HSP90α which was used as the antigen to generate the HSP90α antibody. All five clones in category C in Figure 14, which contained this part of HSP90α, were detected by the HSP90α antibody in the second screening, while the other six clones were not detected. In summary, this indicates polypeptide synthesis from translation start sites within the transcribed inserts recognised by *E. coli* ribosomes.

Antibody cross-reactivity

Screenings with the GAPDH and HSP90α antibodies detected a number of non-GAPDH or HSP90α clones. These false positives reflect a limited specificity of the antibodies. This is not surprising as antibody specificity is determined by binding affinity to certain epitopes and is not usually tested against a whole library of over-expressed proteins. A high number of non-GAPDH clones was recognised by the polyclonal GAPDH antibody but gave relatively weak signals which were easily distinguished from the strong specific signals. In all screening experiments with the monoclonal HSP90α antibody, very little background was observed and essentially all signals could be scored as clear positives. Three non-HSP90α clones were detected by the HSP90α antibody, two of them were identical, and the cross-reactivity of the antibody was confirmed in western blots.

This suggests an interesting application of the described technology for screening antibodies against arrays of potential antigens to detect common epitopes on different proteins. Protein expression libraries on high-density filters can be used for screening the specificity of antibodies against large numbers of antigens arrayed on a solid support. For example, antibodies with no known antigen specificity (e.g. lymphoma proteins) can be screened for binding to a highly diverse repertoire of protein molecules. As all of these proteins are expressed from defined clones of a cDNA library, the corresponding cDNA can easily be sequenced to identify the encoding gene. Such an approach could be applied to homology studies on protein families, defining binding domains, epitopes and interacting molecular motifs. The technique is not limited to antigen-antibody screening but could also be used for other protein-protein interactions or ligand-receptor systems, including non-protein molecules.

6.4 Characterisation of expression products

For the generation of a gene product catalogue in the form of a collection of expression clones, it is necessary to characterise the expression products of large numbers of clones with regard to of protein size, expression strength, solubility and homogeneity. Different techniques are available for protein analysis which may be suitable for high-throughput experimentation. SDS-PAGE was used for the characterisation of clones of the hEx1 library, which is a well established and robust method that is used to separate protein mixtures or analyse isolated proteins. On the other hand, SDS-PAGE involves time-consuming manual steps for preparation of gels, staining and image taking. Capillary electrophoresis (CE) appears advantageous not only in this respect. Automatic sample loading, analyte detection and quantification are available for this technique. To allow high-throughput experimentation, instruments with 96 capillaries have recently been made available by commercial suppliers, and are also being developed at the MPI-MG in the group of C. Heller. CE can be used with native proteins, but this usually has to be optimised to meet the requirements of the individual proteins (122). Isoelectric focusing by CE is a powerful technique (123), but of course no information on protein size is obtained. Analogous to SDS-PAGE, CE has been used with SDS-containing buffers to separate SDS-protein complexes (124). This technique, in combination with a multiple-capillary electrophoresis device, might become a valuable tool for the characterisation of large numbers of expression clones.

Mass spectrometry of proteins and peptides has become a widely used technique. This was mainly enabled by the development of new sample ionisation techniques, as matrix assisted laser desorption/ionisation (MALDI, ref. 86) and electrospray ionisation (125). In this study, MALDI mass spectrometry was used to characterise His_6-tag fusion proteins expressed in *E. coli*. The masses of peptides in tryptic digests of denatured proteins were measured. For sample preparation, His_6-tag fusion proteins were immobilised on magnetic beads, which allowed washing, buffer changes and digestion in a 96-well microtitre plate format. Because of the considerable accuracy of the method, the identity of the proteins was verified with high confidence. Consistence of the observed and the predicted tryptic digestion products was confirmed for large parts of the proteins, but not all of the expected peptides were found in the spectra. Missing peptides may be due to incomplete digestion or inefficient desorption of certain peptides in the MALDI process. Since generally it cannot be ruled out that predicted peptides are missing because the sequence of the expressed protein differs from the DNA

may be the translational apparatus of bacterial cells. Bacteria have polycistronic transcripts, and translation initiates at any ribosome binding site that is followed by a start codon. Eukaryotic cDNA sequences may contain such sites by chance, leading to initiation of translation of the marker protein within the cDNA insert. In contrast, eukaryotic transcripts are monocistronic, and translation is exclusively initiated at a start codon in proximity to the 5'-end of transcripts. Translation cannot start from internal start codons, which might permit the establishment of open reading frame vectors in eukaryotic expression hosts like yeast.

sequence, the use of tryptic digests analysed by MALDI-MS to verify the expression of predicted protein sequences is limited. The coverage of the whole protein sequence may be improved by including additional proteases or cyanogen bromide to complement the information obtained from the tryptic digests. Alternatively, mass spectrometry may be used to determine the mass of the complete expression product, even though the performance of MALDI mass spectrometry of whole proteins may be more variable than for peptide mixtures.

6.5 Perspectives

For the future, further characterisation of the hEx1 library, and the construction of new expression libraries are envisaged.

Characterisation of the hEx1 library

By oligonucleotide fingerprinting of the hEx1 library, clones of identical genes could be grouped into clusters, and the number of different transcripts represented in the library could be determined. Clones would then be identified by comparison with oligonucleotide fingerprints of sequenced IMAGE clones (78), that are currently being generated at the MPI-MG, or by DNA sequencing of one clone per cluster.

The characterisation of the expression products of significant numbers of clones in the hEx1 library requires high-throughput techniques. Growth of bacteria, protein expression and purification in microtitre plates has already been established. To analyse the expression products, different options exist including SDS-PAGE, capillary electrophoresis and mass spectrometry.

The identification of clones in the hEx1 library and the characterisation of the expression products of these clones will provide a catalogue of expression products for a large number of genes. This catalogue should be a valuable tool for the elucidation of the three dimensional structures of large numbers of proteins which has recently been proposed under the term structural genomics (126,127). The elucidation of hundreds of human protein structures is the goal of the ‘Leitprojekt Protein-Strukturfabrik’ (http://userpage.chemie.fu-berlin.de/~ifvsb/), a collaboration of universities, research institutes and companies in Berlin.

Construction of new libraries

For the construction of new expression libraries, alternative expression hosts, cDNA construction techniques and protein purification tags can be envisaged. Oligo(dT)-primed

cDNA is biased towards 3'-ends of genes which leads to missing N-terminal parts of larger proteins. In order to include a maximum number of epitopes for antibody screening, complementary random-primed cDNA libraries will be constructed.

Suitable peptide tags fused to expression products have to meet two criteria. In the first place, the tag has to allow protein purification by affinity chromatography. In this regard the His_6-tag is advantageous, because proteins can be purified in their native state or, if inclusion bodies are formed, under denaturing conditions. The second criterion is the selective (e.g. immunological) detection of the tag on high-density protein filters. An antibody against the tag peptide should be available, that does not cross-react with intrinsic antigens of the host. The tag should be short and therefore unstable when expressed without a protein fused to it. This would allow the detection of expression clones by screening for tag expression.

A number of binding epitopes of monoclonal antibodies have been introduced as tags. These tags are around ten amino acids long, and allow highly specific detection of the expression product in protein mixtures. The use of an epitope tag for purification of fusion proteins is limited, if elution buffers of high or low pH are used. The nine amino acid StrepII-tag allows efficient affinity chromatography on a matrix with modified streptavidin (25,26). This tag can be detected with streptavidin conjugates on blots, but biotinylated host proteins may also be detected. As an alternative to the His_6-tag, the StrepII-tag could be combined with a antibody epitope tag to combine highly specific detection and affinity purification under mild conditions.

Only a small part of random clones in expression libraries express recombinant protein, therefore large numbers of clones have to be picked in order to include expression clones of rare transcripts. It would therefore be advantageous to selectively grow clones that express their cDNA inserts in the correct reading frame. Open reading frame vectors that would allow for such a selection, could be constructed by fusion of a selectable marker to the C-terminus of the cDNA insert. Expression of the marker should only occur if the cloned insert is translated. For expression of the marker, the insert has to contain an open reading frame without stop codons, which is fused in frame to the translation start encoded by the vector and to the marker sequence. This implies that cDNA generated by random priming has to be used, as oligo(dT) primed cDNA contains open reading frames terminated by stop codons. The construction of an open reading frame vector for *E. coli* has been attempted (ref. 128, A. Lüking, K. Büssow, unpublished observations). Howerer, it was observed that marker proteins were expressed, even if the fusion with the cDNA insert was out of frame. The reason for this