Genomics **65,** 1–8 (2000)
doi:10.1006/geno.2000.6141, available online at http://www.idealibrary.com on IDEAL®

A Human cDNA Library for High-Throughput Protein Expression Screening

Konrad Büssow,[1] Eckhard Nordhoff, Christine Lübbert, Hans Lehrach, and Gerald Walter

Max Planck Institute of Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Received October 27, 1999; accepted January 21, 2000

We have constructed a human fetal brain cDNA library in an *Escherichia coli* expression vector for high-throughput screening of recombinant human proteins. Using robot technology, the library was arrayed in microtiter plates and gridded onto high-density filter membranes. Putative expression clones were detected on the filters using an antibody against the N-terminal sequence RGS-His$_6$ of fusion proteins. Positive clones were rearrayed into a new sublibrary, and 96 randomly chosen clones were analyzed. Expression products were analyzed by SDS–PAGE, affinity purification, matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry, and the determined protein masses were compared to masses predicted from DNA sequencing data. It was found that 66% of these clones contained inserts in a correct reading frame. Sixty-four percent of the correct reading frame clones comprised the complete coding sequence of a human protein. High-throughput microtiter plate methods were developed for protein expression, extraction, purification, and mass spectrometric analyses. An enzyme assay for glyceraldehyde-3-phosphate dehydrogenase activity in native extracts was adapted to the microtiter plate format. Our data indicate that high-throughput screening of an arrayed protein expression library is an economical way of generating large numbers of clones producing recombinant human proteins for structural and functional analyses.

INTRODUCTION

Cellular functions are controlled by the networked expression of gene catalogues. Functional network analysis requires the parallel expression and characterization of large numbers of gene products. Structural analysis provides clues to biochemical functions of unknown proteins (Hwang *et al.*, 1999; Zarembinski *et al.*, 1998). Genome analysis by DNA hybridization and sequencing has become a highly automated process (Lehrach *et al.*, 1997). In contrast, the individuality of protein molecules demands highly customized procedures for their expression. Automation of these procedures requires systems that allow the efficient handling of large numbers of clones representing many different proteins. Bacterial systems are easy to manage but the expression of eukaryotic proteins can be problematic, due to aggregation, formation of insoluble inclusion bodies, and/or degradation of the expression product (Hockney, 1994; Makrides, 1996). Eukaryotic systems suffer from lower yields of heterologous protein (e.g., *Saccharomyes cerevisiae;* Buckholz and Gleeson, 1991), high demands on sterility (e.g., mammalian systems; Aruffo, 1997; Kingston *et al.*, 1997), or time-consuming cloning procedures (e.g., Baculovirus system; Miller, 1993).

We have shown that automated technology can be used for high-throughput protein expression screening (Büssow *et al.*, 1998; Lueking *et al.*, 1999). Mammalian cDNA libraries are directly cloned into bacterial expression vectors, circumventing the subcloning of individual protein-coding sequences. In a first screening round, putative protein-expressing clones are identified on high-density filters using antibodies against a vector-encoded tag sequence. The detected clones are rearrayed into a smaller sublibrary. In a second round, small-scale protein expression is performed in microtiter plates. Products are analyzed for size, yield, homogeneity, and solubility using SDS–PAGE, affinity purification, and matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry (MALDI–TOF–MS). Expression levels of large numbers of clones are assessed in parallel to find the most suitable for high-throughput structural analyses by X-ray crystallography or NMR and functional screening. In a third round, protein function is also assayed in the microtiter plate format. As an example, bacterial lysates of 96 clones were screened for expression of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) activity. In summary, our multistep screening approach enables the generation of an expression clone catalogue of human proteins as a resource for structural and functional genomic analyses.

[1] To whom correspondence should be addressed. Telephone: +49-30-8413-1614. Fax: +49-30-8413-1128. E-mail: buessow@molgen.mpg.de.

0888-7543/00 $35.00

MATERIALS AND METHODS

cDNA library and protein expression screening on high-density filters. A cDNA library (hEx1) from human fetal brain tissues was cloned in the expression vector pQE30NST (GenBank Accession No. AF074376). High-density protein filters were prepared and were screened with the RGS · His antibody (Qiagen), as described (Büssow *et al.*, 1998). In total, 193,536 clones of the hEx1 library were picked and labeled according to the RZPD nomenclature (http://www.rzpd.de). Clone names contain the library number MPMGp800 as a prefix.

PCR amplification and sequencing. cDNA inserts were amplified using primers pQE65 (TGAGCGGATA ACAATTTCAC ACAG) and pQE276 (GGCAACCGAG CGTTCTGAAC) at an annealing temperature of 65°C. PCR products were tag-sequenced using primer pQE65.

Protein expression and nickel chelate affinity chromatography in microtiter plates. Proteins were expressed in 1-ml cultures in deep-well microtiter plates, and protein extracts were obtained as described (Lueking *et al.*, 1999). Twenty-five microliters of 50% Ni-NTA agarose was added to protein extracts obtained under denaturing conditions, and His_6-tagged proteins were bound by shaking for 1 h in a microtiter plate shaker. The agarose beads were washed three times by resuspending them in Buffer C (8 M urea, 0.1 M NaH_2PO_4, 0.01 M Tris, pH 6.3), shaking for 5 min, and removal of liquid on the vacuum filtration manifold. Buffer C was removed by washing four times with 200 μl of 5 mM Tris–HCl, pH 8.0. Proteins were eluted by adding 100 μl of 35% acetonitrile, 0.1% TFA, shaking for 10 min, followed by centrifugation at 2000 rpm for 2 min, and collection of eluates in a fresh 96-well microtiter plate. Five microliters of the eluates was analyzed by SDS–PAGE, and 0.5-μl aliquots were subjected to MALDI–TOF-MS analysis.

MALDI–TOF-MS analyses. Aliquots (0.5 μl) of protein eluates were loaded onto a Bruker Scout-384 MALDI sample support (384 sample positions arranged according to the microtiter plate format), followed by addition of 0.5 μl sinapic acid matrix solution (saturated in 35% acetonitrile). The samples were deposited onto the central positions E7–E18. . .L7–L18. In addition, a protein calibration standard containing 0.5 pmol horse heart cytochrome c and 1 pmol human carbonic anhydrase was placed between the sample positions H12, I12, H13, and I13. All samples were analyzed on a Bruker Scout 384 Biflex III MALDI–TOF mass spectrometer in linear operational mode using externally determined calibration constants. Exclusively positively charged ions were detected, and 100–150 single-shot spectra were accumulated for improved signal-to-noise ratio. Before the analysis, the instrumental parameters were optimized for good signal resolution in the mass range 10–30 kDa using the protein calibration standard, and external calibration constants were determined using the molecular ion signals of cytochrome c and human carbonic anhydrase I. If indicated by the above measurements, proteins contained in 10 μl eluate were neutralized and reduced by addition of 2 μl containing 500 mM Tris–HCl, pH 7.5, 50 mM DTT and incubated at 55°C for 30 min. Aliquots (0.5 μl) of these solutions were deposited onto the MALDI sample support followed by 0.5 μl sinapic acid matrix solution containing also 2.5% TFA. After solvent evaporation, these samples were analyzed as described above.

GAPDH assay. The GAPDH assay described by Heinz and Freimüller (1982) was adapted to the microtiter format and performed in duplicate. One hundred fifty microliters of assay mix (33 mM TEA–HCl, 0.23 mM NADH, 6.7 mM $MgSO_4$, 1 mM ATP, 3 mM glycerate 3-phosphate, 3.8 mM L-cysteine) was added to 1-μl soluble protein fractions diluted 1:10, in 96-well microtiter plates (Microtest III, Falcon). The decrease of A_{340} was measured with a microtiter plate photometer (Spectramax 250, Molecular Devices). One unit of GAPDH catalyzes the reduction of 1 μmol of 1,3-diphosphateglycerate to D-glyceraldehyde-3-phosphate per minute.

RESULTS

A human fetal brain cDNA expression library (hEx1) was constructed in the vector pQE30NST, which allows the expression of fusion proteins with the N-terminal sequence RGS-His_6 (Büssow *et al.*, 1998). Briefly, 193,536 clones were picked into 384-well microtiter plates (plates 1 to 504 of hEx1) and gridded as high-density protein filters using a robotic system (Lehrach *et al.*, 1997). These clone arrays were screened for putative protein expression clones using the monoclonal antibody RGS · His (Qiagen), which recognizes the N-terminal sequence RGS-His_6 of recombinant expression products. The antibody preferentially labels clones containing a cDNA insert in-frame with RGS-His_6. In alternative reading frames, stop codons cause the expression of short and unstable products that are degraded in the *Escherichia coli* host cell (Gottesman, 1996). A total of 37,830 (19.6%) clones were recognized by the RGS · His antibody, 67% of which were labeled with medium or high intensity. All positive clones were combined in a new library by rearraying in 99 × 384-well microtiter plates (labeled plates 505 to 603 of hEx1) using the same robotic system equipped with dedicated rearraying software (Büssow *et al.*, 1998).

cDNA inserts of 96 randomly chosen clones of the medium and high RGS · His signal intensity groups were sequenced. All 96 clones originated from plate 582 of the rearrayed hEx1 library. cDNA inserts were amplified by PCR and analyzed by 5′-tag sequencing. An average insert size of 1.5 kb was determined. 5′-tag sequences of 93 cDNA inserts were obtained and used to search SP-TrEMBL, the combined SWISS-PROT and TrEMBL protein database (Bairoch and Apweiler, 1998) using the program BLASTX (Altschul *et al.*, 1990). Fifty-nine sequences were found to match human proteins in this database (Table 1). Thirty-eight (64%) of those sequences matched the beginning of a human protein, suggesting that the complete coding region had been cloned (full-length clones). Thirty-nine (66%) of the 59 sequences were fused to the N-terminal sequence RGS-His_6 in the correct reading frame (RF+). Protein molecular masses were predicted for these clones by completing their 5′-tag sequences using the matching sequences in the database (Table 1), considering that the formyl group of the N-terminal formylmethionine is removed in *E. coli* and that the resulting N-terminal methionine is usually not removed if it is followed by arginine (Sherman *et al.*, 1985).

Expression products of the same 96 clones were analyzed by SDS–PAGE of cellular protein extracts and by nickel chelate affinity purification, followed by both SDS–PAGE and MALDI–TOF-MS. Protein expression and all subsequent steps were performed in 96-well microtiter plates. Seventy-two (75%) of the total 96 clones expressed recombinant proteins detectable in SDS–PAGE. Thirty-five of the 39 in-frame clones produced RGS-His_6 tag fusion proteins of expected sizes,

TABLE 1

Protein Expression Properties of hEx1 Clones with Sequence Database Matches

Clone MPMGp 800. . .	SP-TrEMBL database match		First matched amino acid in database sequence	Reading frame (RF)	Predicted protein size (kDa)	Expressed protein size, measured by MALDI-TOF-MS (kDa)	Expressed protein size, estimated by SDS–PAGE (kDa)[a]
	Accession No.	Protein name					
A06582	O00217	NADH-ubiquinone oxidoreductase 23 kDa subunit precursor (EC 1653)	1	−	—	8,602	12
A08582	P04687	Tubulin α-1 chain	174	+	33,859	33,865	35
A12582	P04765	Eukaryotic initiation factor 4A-I (EIF-4A-I)	12	+	47,808	47,819	50
A14582	Q11203	CMP-*N*-acetylneuraminate-B-1,4-galactoside α-2,3-sialyltransferase	234	+	18,825	18,831[b]	20
A16582	P13639	Elongation factor 2 (EF-2)	1	+	98,098	(18,458)	(97)
A18582	P49006	Marcks-related protein (MAC-MARCKS)	1	−	—	12,282	14
A20582	P56182	NNP-1 protein (D21S2056E)	240	+	27,850	27,857	34
A24582	P49006	Marcks-related protein (MAC-MARCKS)	1	−	—	10,052	13
C06582	Q15853	Upstream stimulatory factor 2	128	−	—	13,908	18
C10582	P36578	60S Ribosomal protein L1 (L4)	1	−	—	10,304	10
C12582	P49006	Marcks-related protein (MAC-MARCKS)	1	−	—	13,557	16
E02582	P43308	Translocon-associated protein, β subunit precursor (TRAP-β)	1	−	—	10,753	10
E04582	P14793	60S Ribosomal protein L40 (CEP52)	36	−	—	7,750	10
E10582	O75312	Zinc-finger protein ZPR1	1	+	54,299	54,334	55
E12582	P25111	40S Ribosomal protein S25	1	−	—	7,504	10
E14582	P04687	Tubulin α-1 chain	304	−	—	11,301	10
E18582	Q15560	Transcription elongation factor S-II	1	+	39,238	8,445	10
E20582	Q13885	β Tubulin	275	+	22,579	22,594	23
G02582	P14923	Junction plakoglobin	287	+	53,034	53,066	50
G04582	Q16478	Glutamate receptor subunit	832	−	—	12,653	n.e.
G10582	P07108	Acyl-CoA-binding protein (ACBP)	1	+	15,098	15,109[c]	15
G12582	P48735	Isocitrate dehydrogenase (NADP), mitochondrial precursor (EC 11142)	1	+	55,802	55,832	50
G14582	P39023	60S Ribosomal protein L3	225	−	—	8,904	10
G16582	P15880	40S Ribosomal protein S2 (S4)	1	+	34,328	34,302	35
G20582	P54198	HIRA protein	383	+	72,278	72,316[c]	65
I02582	P36404	ADP-ribosylation factor-like protein 2	1	−	—	12,703	12
I04582	P30086	Phosphatidylethanolamine-binding protein	1	+	27,046	27,029[c]	29
I06582	P25111	40S Ribosomal protein S25	1	+	17,678	17,685	23
I10582	P39023	60S Ribosomal protein L3	1	+	49,026	49,037[c]	50
I12582	P15880	40S Ribosomal protein S2 (S4)	1	+	34,328	34,329[c]	35
I14582	Q13098	G Protein pathway suppressor 1 (GPS1 protein)	193	−	—	17,996	18
I18582	P05092	Peptidyl-prolyl *cis*-trans isomerase A	1	+	21,441	21,460	23
I20582	P02570	Actin, cytoplasmic 1 (β-actin)	1	+	47,297	47,338	45
I24582	P23396	40S Ribosomal protein S3	1	+	29,749	29,761[c]	32
K04582	Q03827	Transcription factor ETR101	97	+	16,184	16,182	23
K08582	Q00403	Transcription initiation factor IIB (TFIIB)	1	+	38,133	38,156	38
K10582	O15143	ARP2/3 complex 41 kDa subunit (P41-ARC)	1	+	46,403	46,405	45
K12582	Q15666	Asparagine synthetase (fragment)	1	−	—	17,806	21
K14582	P49241	40S Ribosomal protein S3A	1	+	33,094	33,095[c]	35
K16582	Q99719	Cell division control-related protein	1	−	—	8,509	12
K18582	P04687	Tubulin α-1 chain	306	+	19,199	19,202	21
K20582	O00240	Dihydropyrimidinase-related protein-4 (DRP-4)	1	−	—	15,257	14
M02582	Q13885	β-Tubulin	253	−	—	26,301	27
M04582	P49368	T-complex protein 1, γ subunit (TCP-1-γ)	19	+	61,326	61,352	60
M10582	P02571	Actin, cytoplasmic 2 (γ-actin)	1	+	46,718	46,749	45
M12582	P32969	60S Ribosomal protein L9	1	−	—	9,950	14
M18582	Q13885	β Tubulin	1	+	54,291	54,307	58
M20582	P02768	Serum albumin precursor	116	+	59,076	59,116	60
M22582	P02570	Actin, cytoplasmic 1 (β-ACTIN)	1	+	47,297	47,306	45
M24582	Q06830	Thioredoxin peroxidase 2	1	+	26,231	26,222	25
O02582	Q02878	60S Ribosomal protein L6	1	+	35,457	35,460[c]	35
O04582	Q02878	60S Ribosomal protein L6	1	+	35,457	35,467[c]	35
O06582	P17080	GTP-binding nuclear protein RAN (TC4)	1	+	28,494	28,516[c]	30

TABLE 1—*Continued*

Clone MPMGp 800. . .	SP-TrEMBL database match: Accession No.	SP-TrEMBL database match: Protein name	First matched amino acid in database sequence	Reading frame (RF)	Predicted protein size (kDa)	Expressed protein size, measured by MALDI-TOF-MS (kDa)	Expressed protein size, estimated by SDS–PAGE (kDa)[a]
O08582	Q08379	Golgin-95	414	+	26,312	26,303	30
O10582	P01922	Hemoglobin α-chain	1	+	15,126	15,122	20
O14582	P21810	Bone/cartilage proteoglycan I precursor (biglycan) (PG-S1)	1	−	—	9,231	14
O16582	Q15597	Translation INITIATION FACTOR EIF-4γ (fragment)	215	+	58,069	58,108	60
O18582	Q14257	Calcium-binding protein ERC-55 precursor	24	+	42,157	42,165	60
O20582	Q02543	60S Ribosomal protein L18A	1	+	24,374	24,379[c]	26

Note. RF ± reading frame of insert in relation to His_6-tag. Expressed protein size was determined by MALDI-TOF-MS and estimated by SDS–PAGE referring to the band of the largest size visible against the *E. coli* background.

[a] n.e., no expression observed.

[b] Prior to reduction with DTT, abundant signals corresponding to the monomer, monomer + 1 glutathione residue, and the protein dimer were observed.

[c] Prior to reduction with DTT the determined mass was 295–315 Da higher (+ glutathione).

while in the remaining four clones (E18582, K10582, O02582, and O04582) no expression products could be detected (Table 1). His_6-tagged proteins were affinity-purified under denaturing conditions using Ni-NTA agarose beads and filter plates (Fig. 1). Six expression products, including E18582, K10582, O02582, and O04582, which were not detected in whole cellular protein extracts, could be identified after purification. Protein sizes were determined by SDS–PAGE and MALDI–TOF-MS (Fig. 2), and both data sets were

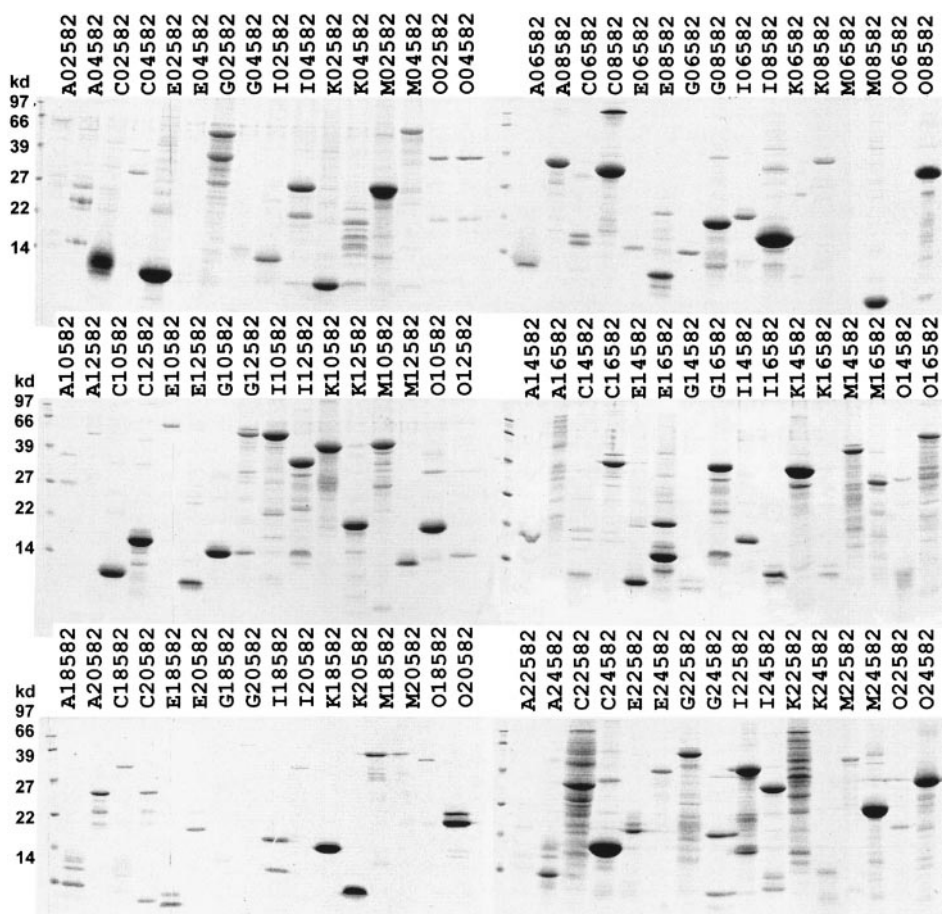

FIG. 1. SDS–PAGE of nickel chelate purified proteins. Following protein expression in microtiter plates, cells were lysed under denaturing conditions. His_6-tagged proteins were purified by nickel chelate chromatography in microtiter plates and were analyzed by SDS–PAGE, followed by Coomassie staining. Lanes are labeled using RZPD clone names, omitting the prefix MPMGp800.

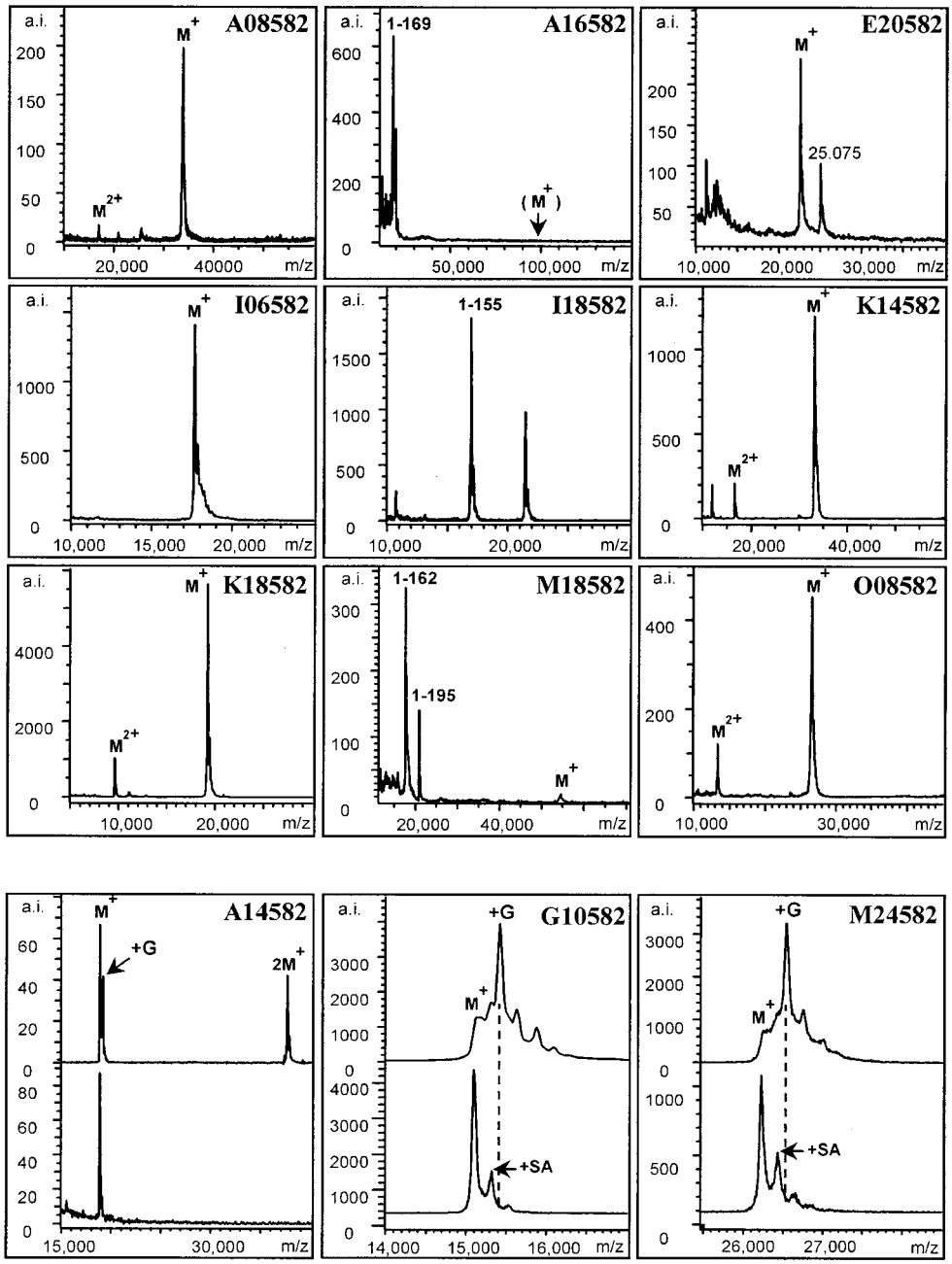

FIG. 2. MALDI–TOF-MS of nickel chelate purified proteins. Following nickel chelate affinity purification, the obtained eluates were analyzed by MALDI–TOF-MS. (**Top three panels**) Some recorded mass spectra; clones and labeling as in Fig. 1. M^+ and M^{2+}, singly and doubly charged molecular ions of expected expression products. The numbers of the first and last amino acids indicate assigned C-terminally truncated protein sequences. For clone A16582, the expected protein (98,098 kDa) could not be detected. For some expression products, the determined molecular masses exceeded the predicted values by approximately 300 Da (Table 1), indicative of glutathionylation. In addition, for A14582 a strong protein-dimer molecular ion signal was recorded indicative of protein–protein disulfide bridges. These indications were verified by reduction with DTT prior to the mass spectrometric analysis. (**Bottom**) Mass spectra obtained from A14582, G10582, and M24582 before (**top spectrum**) and after (**bottom spectrum**) reduction with DTT. +G, single glutathionylation; +SA, adduction of one sinapic acid molecule used as MALDI matrix.

compared to the corresponding values predicted from DNA sequencing data (Table 1).

As expected, the predicted molecular masses were considerably better matched by the masses determined with MALDI–TOF-MS than by those estimated from SDS–PAGE (Table 1). For most clones, the determined molecular mass deviated less than 0.1% from the predicted value. The fusion protein expected for clone A16582 (98.098 kDa) could not be detected; instead an abundant signal at *m/z* 18,446 dominated the recorded spectrum. Considering that the N-terminal sequence RGS-His_6 is vital for the applied affinity purification and that N-terminal methionine is usually not removed within *E. coli* if followed by arginine (Sherman

et al., 1985), this signal could be assigned to the truncated sequence 1–169 (expected *m/z* 18,442). SDS–PAGE of the purified A16582 protein showed a number of bands ranging from approximately 20 to approximately 100 kDa (Fig. 1). These results suggest that the protein is unstable in *E. coli*. An incomplete expression product was also observed for clone E18582. Various other clones produced abundant C-terminally truncated expression products in addition to the expected product. Figure 2 shows a selection of the recorded mass spectra including signal interpretation. Clone E20582 expressed a 25,075-Da protein of unknown identity in addition to the expected 22,579-Da protein. A possible explanation would be a frameshift mutation leading to a larger expression product in a subpopulation of the E20582 *E. coli* cells.

For the clones A14582, G10582, G12582, G16582, G20582, I04582, I10582, I12582, I24582, K04582, K10582, M24582, O02582, O04582, and O06582, the determined molecular masses exceeded the expected values by 0.290–0.320 kDa (three examples are shown in Fig. 2). This deviation is indicative of glutathionylation (attachment of one glutathione residue to a cysteine residue by formation of a disulfide bridge), an essential intermediate reaction of disulfide bridge reduction in *E. coli*. In addition, for the clone A14582 a strong signal corresponding to the molecular mass of the protein dimer was detected (Fig. 2). In addition to singly charged molecular ions, to a lower degree nonspecific multimers as well as doubly and triply charged molecular ions are formed during MALDI–TOF-MS of protein samples. Therefore, peak intensities must be taken into consideration to recognize protein dimers in the sample. Since the MALDI sample preparation conditions used (pH <2, 35% acetonitrile) denature most protein–protein interactions, a covalent linkage is likely to account for the observed dimers. Both glutathionylation and protein–protein disulfide bridges were verified by reduction with DTT prior to the mass spectrometric analysis. After reduction, the determined protein masses matched the expected masses in all cases within 0.1% maximum deviation, and no more protein dimers were detected (Fig. 2).

Expression products from inserts cloned in incorrect reading frames (RF−) were generally smaller than those from inserts in the correct reading frame (RF +, Table 1). As shown in Fig. 3, the molecular mass of expression products is correlated to the reading frame of the cDNA insert. Sixteen of 17 expression product smaller than 15 kDa derived from RF− clones, while 31 of 32 expression products of at least 20 kDa size derived from RF+ clones. Thus the molecular mass of a clone's expression product can be used as a measure to predict the reading frame of its cDNA insert, if the DNA sequence is unknown.

The screening of commonly available cDNA expression libraries for functional activities is complicated by large numbers of clones that do not express their cDNA inserts as proteins. By arraying, antibody screening,

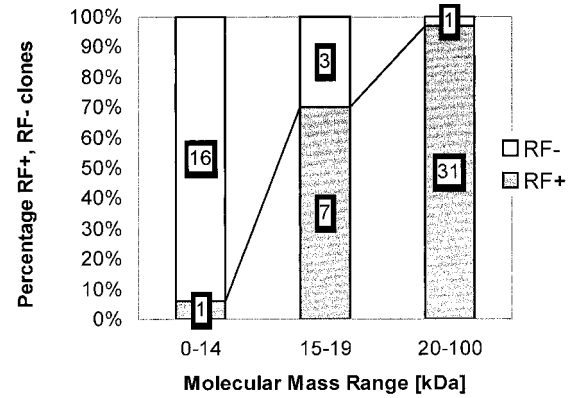

FIG. 3. Expression product size and reading frame. Relationship between the size of expressed recombinant protein and reading frame in the same clones as in Table 1. Clones represented by gray bars contain cDNA inserts translated in the correct reading frame (RF+), whereas in the other clones (white bars) translation can occur only in an incorrect reading frame (RF−). Numbers of clones are indicated in the bars.

and molecular mass detection, putative expression clones can be detected with a high level of efficiency. Therefore smaller numbers of clones must be assayed, and functional screening in microtiter plates becomes practicable. As an example, a GAPDH activity assay was adapted to the microtiter plate format. A positive control clone (D215) expressing human GAPDH as an RGS–His_6 tag fusion protein was introduced in exchange for one of the 96 hEx1 clones of Fig. 1. Protein expression was induced, and cells were lysed in 150 μl lysis buffer under nondenaturing conditions. GAPDH activities were measured in 0.1-μl aliquots of the lysates, and the positive control clone (D215) was clearly identified (Fig. 4). Duplicate experiments gave identical activity patterns with at least three additional clones (C04582, C06582, and C08582) above an arbitrary background. Two of these clones express products that did not match human proteins in the database, while the third represents a short out-of-frame fragment.

DISCUSSION

Structural genomics is expected to provide a link between DNA sequence information and protein function (Gaasterland, 1998; Kim, 1998; Rost, 1998). This requires the expression and characterization of large numbers of human proteins. We have shown that desired protein expression clones can be selected at high throughput from an arrayed cDNA library using a multistep screening procedure. This highly parallel approach seems to be an efficient alternative to the subcloning of individual cDNA sequences. In a first step, high-density protein filters are screened for putative expression clones (Büssow *et al.*, 1998; Lueking *et al.*, 1999). Those clones are then rearrayed into a sublibrary enriched for in-frame inserts. DNA sequence analysis of 93 randomly chosen clones of the hEx1

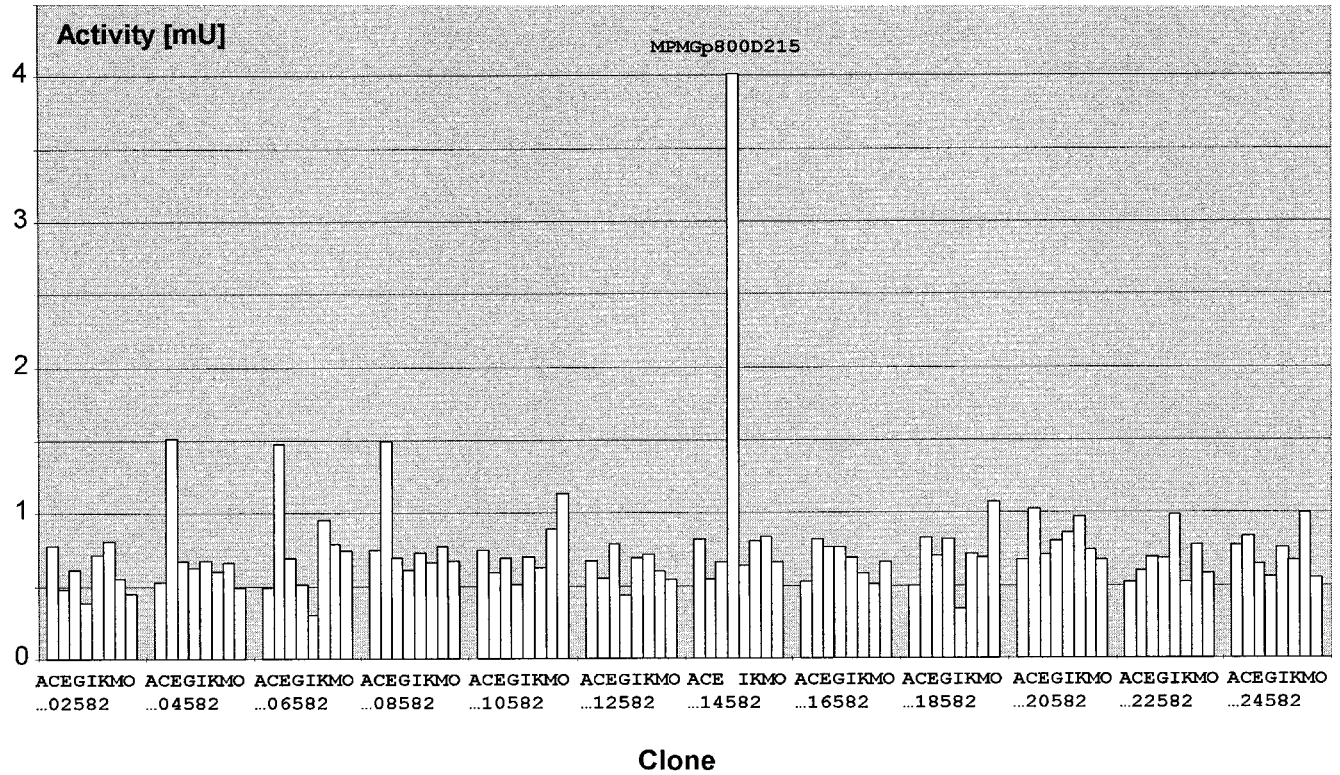

FIG. 4. GAPDH activity screen. Screening for GAPDH activity in bacterial lysates of 96 clones of the hEx1 library; clone numbers and labeling as in Fig. 1; positive control, MPMGp800D215 containing a GAPDH cDNA insert.

library showed that among the known genes, two-thirds (66%) expressed their inserts in the correct reading frame, reflecting the general efficiency of the screening method. The two-thirds (64%) of clones containing the complete coding sequence of a human protein do of course reflect the usual bias toward smaller products of cDNA libraries (Table 1).

In a second step, proteins are expressed in microtiter plates, and SDS–PAGE, nickel chelate affinity purification, and MALDI–TOF-MS are used to identify the best protein-expressing clones. DNA sequencing was used to correlate reading frames with protein sizes determined by SDS–PAGE and MALDI–TOF-MS. Although both methods returned compatible results, the latter was found to give the most accurate indication for the reading frame of cDNA inserts, because nearly all clones with expression products of at least 20 kDa had inserts in the correct reading frame (Fig. 3). This indicates that MALDI–TOF-MS size selection is a useful criterion for confirmation of expression clones. The molecular masses determined by MALDI–TOF-MS are in good agreement with the values predicted from the corresponding DNA sequences (Table 1). This reflects the expected experimental mass accuracy considering that all samples were analyzed using identical instrumental settings and the same external calibration constants. These examples demonstrate the power of MALDI–TOF-MS for characterizing cDNA expression products at high throughput. The detection is sensitive (midfemtomolar range), is rapid ($<$1 min per sample), and provides detailed and accurate information about the status and purity of the expression products. Questions as to whether a certain clone produces high-quality protein for X-ray or NMR analysis or whether size exclusion chromatography following affinity purification can provide the necessary purity and homogeneity can be addressed early and at low cost. MALDI–TOF-MS analysis of whole libraries will provide a catalogue of expression clones for large numbers of human proteins. By including tryptic peptide mass fingerprinting, new clones will be directly identified by database comparison (Pappin *et al.*, 1993).

In a third step, the microtiter plate technology was extended to functional screening. A spectrophotometric enzyme assay was developed that detects GAPDH expression clones among 96 hEx1 library clones, using nondenaturing bacterial lysates. It is expected that this kind of assay can be adapted to screen expression libraries for other biological activities in the microtiter plate format. Only 1/1000 of the bacterial lysate was used for the GAPDH assay. Therefore, the amount of protein is not expected to be limiting in future assays. Even if the bulk of a protein of interest is expressed in insoluble form, a small soluble fraction could be sufficient for detection. If necessary, affinity purification in microtiter plates can be used to reduce the background of *E. coli* proteins.

Up-scaling of the assay from 96 clones to the whole library will enable the detection of more clones with low to medium activity. This might include a certain degree of false-positive background but will also detect

new biological activity of yet uncharacterized human proteins.

In summary, the use of robot technology for handling and arraying of cDNA libraries, in combination with high-throughput microtiter plate techniques and MALDI–TOF-MS for the analysis of gene products, enables the generation of a catalogue of expression clones as a tool for the characterization of the human proteome.

ACKNOWLEDGMENTS

We thank David Bancroft and Martin Horn for conversion of a picking robot to a rearraying robot, Sabine Thamm for DNA sequencing, Steffen Hennig for sequence database searches, Janett Tischer, Dirk Riebensahm, Sabine Lau, and Claudia Gutjahr for technical assistance, and Angelika Lüking, Caterina Holz, Dolores Cahill, and Eberhard Scherzinger for valuable discussions.

Note added in proof. DNA sequences of hEx1 clones and high-density DNA and protein filters of the rearrayed hEx1 library are publicly available at the Resource Center Primary Database within the German Human Genome Project (http://www.rzpd.de).

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Aruffo, A. (1997). Transient expression of proteins using COS cells. *In* "Current Protocols in Molecular Biology" (F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, *et al.*, Eds.), pp. 16.13.11–16.13.17, Wiley, New York.

Bairoch, A., and Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res.* **26:** 38–42.

Buckholz, R. G., and Gleeson, M. A. (1991). Yeast systems for the commercial production of heterologous proteins. *Bio/Technology* **9:** 1067–1072.

Büssow, K., Cahill, D., Nietfeld, W., Bancroft, D., Scherzinger, E., Lehrach, H., and Walter, G. (1998). A method for global protein expression and antibody screening on high-density filters of an arrayed cDNA library. *Nucleic Acids Res.* **26:** 5007–5008.

Gaasterland, T. (1998). Structural genomics: Bioinformatics in the driver's seat. *Nat. Biotechnol.* **16:** 625–627.

Gottesman, S. (1996). Proteases and their targets in *Escherichia coli*. *Annu. Rev. Genet.* **30:** 465–506.

Gubler, U., and Hoffman, B. J. (1983). A simple and very efficient method for generating cDNA libraries. *Gene* **25:** 263–269.

Heinz, F., and Freimüller, B. (1982). Glyceraldehyde-3-phosphate dehydrogenase from human tissues. *Methods Enzymol.* **89:** 301–305.

Hockney, R. C. (1994). Recent developments in heterologous protein production in *Escherichia coli*. *Trends Biotechnol.* **12:** 456–463.

Hoheisel, J. D., Lennon, G., Zehetner, G., and Lehrach, H. (1991). Use of high coverage reference libraries of *Drosophila melanogaster* for relational data analysis—A step towards mapping and sequencing of the genome. *J. Mol. Biol.* **220:** 903–914.

Hwang, K. Y., Chung, J. H., Kim, S.-H., Han, Y. S., and Cho, Y. (1999). Structure-based identification of a novel NTPase from *Methanococcus jannaschii*. *Nat. Struct. Biol.* **6:** 691–696.

Kim, S. H. (1998). Shining a light on structural genomics. *Nat. Struct. Biol.* **5**(Suppl.): 643–645.

Kingston, R. E., Kaufman, R. J., Bebbington, C. R., and Rolfe, M. R. (1997). Amplification using CHO cell expression vectors. *In* "Current Protocols in Molecular Biology" (F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, *et al.*, Eds.), p. 16.14.11, Wiley, New York.

Lehrach, H., Bancroft, D., and Maier, E. (1997). Robotics, computing, and biology. *Interdisc. Sci. Rev.* **22:** 37–44.

Lueking, A., Horn, M., Eickhoff, H., Büssow, K., Lehrach, H., and Walter, G. (1999). Protein microarrays for gene expression and antibody screening. *Anal. Biochem.* **270:** 103–111.

Maier, E., Bancroft, D. R., and Lehrach, H. (1997). Large-scale library characterization. *In* "Automation Technologies for Genome Characterization" (T. J. Beugelsdijk, Ed.), pp. 65–88, Wiley, New York.

Maier, E., Meier-Ewert, S., Ahmadi, A. R., Curtis, J., and Lehrach, H. (1994). Application of robotic technology to automated sequence fingerprint analysis by oligonucleotide hybridisation. *J. Biotechnol.* **35:** 191–203.

Makrides, S. C. (1996). Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol. Rev.* **60:** 512–538.

Miller, L. K. (1993). Baculoviruses: High-level expression in insect cell. *Curr. Opin. Genet. Dev.* **3:** 97–101.

Pappin, D. J. C., Højrup, P., and Bleasby, A. J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3:** 327–332.

Rost, B. (1998). Marrying structure and genomics. *Structure* **6:** 259–263.

Sherman, F., Stewart, J. W., and Tsunasawa, S. (1985). Methionine or not methionine at the beginning of a protein. *BioEssays* **3:** 27–31.

Zarembinski, T. I., Hung, L. W., Mueller-Dieckmann, H. J., Kim, K. K., Yokota, H., Kim, R., and Kim, S. H. (1998). Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics. *Proc. Natl. Acad. Sci. USA* **95:** 15189–15193.