Konrad Büssow

Arrayed cDNA expression libraries for antibody screening and systematic analysis of gene products

Dissertation

eingereicht im Fachbereich Chemie

der Freien Universität Berlin

1998

1. Gutachter: Prof. Dr. H. Lehrach

2. Gutachter: Prof. Dr. V. Erdmann

Die vorliegende Arbeit wurde in der Zeit von Mai 1995 bis November 1998 in der Abteilung Prof. Dr. H. Lehrach am Max-Planck-Institut für Molekulare Genetik in Berlin durchgeführt.
Ich möchte mich bei Herrn Prof. Lehrach bedanken, der die Entstehung dieser Arbeit in seiner Abteilung in großzügiger Weise unterstützt hat und mir mit Ratschlägen zur Seite stand. Besonderer Dank geht an Herrn Dr. G. Walter, der durch anregende Diskussionen und fördernde Kritik Wesentliches zum Entstehen dieser Arbeit beigetragen hat. Bei Herrn Dr. E. Scherzinger bedanke ich mich für die Einführung in die Methoden der Proteinexpression in *E. coli*, sowie für fördernde Anregungen und Diskussionen. Für verschiedene Anregungen zum Manuskript dieser Arbeit danke ich Dr. A. Lüking und Dr. C. Holz.

Contents

7. Abstract

7.1 Abstract in English

Functional analysis of the proteins expressed in a specific tissue and/or stage of development is an essential step towards understanding biological processes. No technique has yet been available to go directly from sequence information on individual cDNA clones to the protein products. In the approach described here, arrayed cDNA expression libraries from human fetal brain (hEx1) and mouse adult kidney (mKd1) were established for the integration of DNA-based and protein-based experimental data. An *E. coli* plasmid vector (pQE30NST) for the expression of His_6-tag fusion proteins was used to clone cDNA synthesised by oligo(dT) priming. Using automated systems, 27,600 clones of the mKd1 and 193,500 clones of the hEx1 library were picked, grown in microtitre plates and arrayed on membrane filters at a density of 9,216 or 27,648 clones per filter. High-density DNA and protein filters were prepared and used to screen the cDNA libraries with DNA probes and antibodies in parallel. An antibody directed against the N-terminal tag sequence $RGSH_6$ of proteins expressed from pQE30NST (RGS·His antibody, Qiagen) was used to detect expression clones. Approximately 20% of clones in the mKd1 and hEx1 cDNA libraries were detected as putative expression clones by this antibody. 37,830 putative expression clones in the hEx1 library were combined and rearrayed into new microtitre plates. A copy of this sub-library was given to the Resource Centre of the German Human Genome Project (RZPD, http://www.rzpd.de) to generate and distribute high-density DNA and protein filters. Expression products and DNA sequences of 96 randomly selected clones were analysed in detail. Protein expression and purification in microtitre plates was established. 68 out of 96 clones expressed proteins of at least 15 kd size (estimated from SDS-PAGE) which were detected by SDS-PAGE of whole cellular proteins or by metal affinity chromatography. DNA sequences derived from 5'-ends of cDNA inserts were obtained for 93 clones. 58 sequences matched human protein sequences in the SWISS-PROT and TrEMBL databases. 38 (66%) of these 58 sequences contained inserts cloned in the correct reading frame, i.e. the His_6-tag and a protein coding cDNA sequence were fused in frame. 66% of sequences matched to the beginning of a protein database sequence, and the

corresponding clones were therefore considered to contain complete protein-coding regions (full-length clones).

Expression clones for a panel of human genes were identified in the hEx1 library. Nine cDNA probes for BMP-7, calmodulin, COX4, GAPDH, hMSH2, HSP90α, HSP90β, RXRβ and VDAC1 were used for screening by DNA hybridisation. Among the positive clones, putative expression clones were selected that were also detected by the RGS·His antibody on high-density filters. Expression clones for seven genes, but not for BMP-7 and VDAC1, were found with this strategy. GAPDH and calmodulin clones comprising full protein-coding regions expressing soluble His_6-tag fusion proteins were identified. Biological activity of His_6-tag GAPDH and calmodulin fusion proteins was shown with enzyme assays. DNA probes and the RGS·His antibody were used in combination to detect clones expressing GAPDH and HSP90α fusion proteins. In parallel, specific antibodies directed against GAPDH and HSP90α were used to identify expression clones on high-density protein filters. All clones identified by the protein-specific antibodies expressing His_6-tag GAPDH or HSP90α fusion proteins were reliably detected by the RGS·His antibody. The RGS·His detected additional clones, which either contained GAPDH or HSP90α inserts in an incorrect reading frame, or which expressed truncated proteins lacking the epitopes recognised by the protein-specific antibody. Estimated 90% of clones with inserts in an incorrect reading frame were not detected by the RGS·His antibody.

A protocol for the analysis of His_6-tag fusion proteins by matrix assisted laser ionisation/desorption mass spectrometry (MALDI-MS) was established. His_6-fusion proteins were purified under denaturing conditions by immobilisation on Ni^{2+}-chelate-coated magnetic beads, followed by direct digestion with trypsin. The masses of the tryptic peptides were measured and compared to masses predicted from sequences in databases.

Arraying of cDNA expression libraries extends the application of DNA library arrays to protein-based screening techniques, thus creating a direct link between DNA-based and protein-based experimental data. The hEx1 expression library allows for the fast identification of expression clones for specific genes by DNA hybridisation or antibody screening. Proteins expressed from library clones can be directly used for functional assays, or, if inclusion bodies are formed, may be used as antigens to generate antibodies or possibly refolded *in vitro*. Libraries of expression clones displayed on high-density filters, in combination with high-

throughput protein expression and purification in microtitre plates, should be a feasible approach to the construction of gene product catalogues.

7.2 Zusammenfassung in deutscher Sprache

Die funktionelle Analyse der Proteine, die in bestimmten Geweben und/oder Entwicklungsstadien exprimiert werden, ist unverzichtbar für das Verständnis biologischer Vorgänge. Bis heute steht keine Technik zur Verfügung, um direkt von Sequenzinformation individueller cDNA Klone zu den Proteinprodukten zu gelangen. Bei der hier beschriebenen Vorgehensweise wurden geordnete cDNA Expressionsbanken aus humanem Fötalgewebe (hEx1) und Maus Nierengewebe (mKd1) hergestellt, um experimentelle Daten, die auf DNA-Basis oder Proteinbasis gewonnen wurden, integrieren zu können. Ein *E. coli* Plasmidvektor (pQE30NST) für die Expression His_6-markierter Fusionsproteine wurde für die Klonierung von cDNA benutzt, die mit einem oligo(dT) Primer synthetisiert wurde. 27.600 Klone der mKd1 und 193.000 Klone der hEx1 Bank wurden mit Hilfe von Robotern gepickt, in Mikrotiterplatten wachsen gelassen und auf Membranfiltern mit einer Dichte von 9.216 oder 27.648 Klonen pro Filtern angeordnet. Hochdichte-DNA und Proteinfilter wurden auf diese Art hergestellt und für das parallele Screening der cDNA-Banken mit DNA-Sonden und Antikörpern benutzt. Ein Antikörper gegen die N-terminale Sequenz $RGSH_6$ von Proteinen, die mit dem pQE30NST Vektor exprimiert wurden, wurde benutzt, um Expressionsklone zu identifizieren. Ungefähr 20% der Klone der mKd1 und hEx1 cDNA Banken wurden durch diesen Antikörper als mutmaßliche Expressionsklone identifiziert. 37.830 mutmaßliche Expressionsklone der hEx1 cDNA-Bank wurden zusammengefaßt und neu in Mikrotiterplatten angeordnet. Eine Kopie dieser neuen cDNA-Bank wurde dem Ressourcenzentrum im Deutschen Humangenomprojekt (RZPD, http://www.rzpd.de) für die Herstellung und Verteilung von Hoch-Dichte DNA- und Proteinfiltern gegeben. Die Expressionsprodukte und DNA-Sequenzen von 96 zufällig ausgewählten Klonen wurden analysiert. Zu diesem Zweck wurde die Expression und Reinigung von Proteinen in Mikrotiterplatten etabliert. 68 von 96 Klonen exprimierte Proteine von mindestens 15 kd Größe (bestimmt durch SDS-Gelelektrophorese), die nach SDS-Gelelektrophorese von Gesamtzellproteinen oder nach Metall-Affinitätschromatographie nachgewiesen wurden. cDNA-Inserts von 93 Klonen wurden vom 5'-Ende her ansequenziert. 58 dieser Sequenzen entsprachen Proteinsequenzen in der SWISS-PROT und TrEMBL Datenbank. 38 (66%)

dieser 58 Sequenzen enthielten Inserts, die im richtigen Leseraster kloniert waren, d.h. die His_6-Sequenz auf dem Vektor und eine Protein-kodierende cDNA Sequenz waren im richtigen Leseraster fusioniert. 66% war der Anteil der Sequenzen, die mit dem Anfang einer Proteinsequenz übereinstimmten, und von denen deshalb angenommen wurde, daß sie eine vollständige Protein-kodierende Sequenz enthalten („full-length Klone“).

In der hEx1 Bibliothek wurden Expressionsklone für eine Auswahl menschlicher Gene identifiziert. Neun cDNA-Sonden für BMP-7, Calmodulin, COX4, GAPDH, hMSH2, HSP90α, HSP90β, RXRβ und VDAC1 wurden für DNA-Hybridisierungs-Screenings benutzt. Unter den positiven Klonen wurden solche Klone, die auch durch den RGS·His-Antikörper detektiert wurden, als mutmaßliche Expressionsklone betrachtet. Expressionsklone für sieben Gene, jedoch nicht für BMP-7 und VDAC1, wurden so gefunden. Biologische Aktivität von His_6-GAPDH und Calmodulin Fusionsproteinen wurde durch Enzym-Assays nachgewiesen. DNA-Sonden und der RGS·His Antikörper wurden in Kombination eingesetzt, um Klone zu identifizieren, die GAPDH und HSP90α Fusionsproteine exprimieren. Parallel dazu wurden spezifische Antikörper gegen GAPDH und HSP90α eingesetzt, um Expressionsklone auf Hochdichte-Proteinfiltern zu identifizieren. Alle Klone, die von den Protein-spezifischen Antikörper erkannt wurden und His_6-GAPDH oder HSP90α Fusionsproteine exprimierten, wurde auch von dem RGS·His Antikörpern erkannt. Der RGS·His Antikörper detektierte weitere Klone, die GAPDH oder HSP90α Inserts enthielten, die entweder nicht im richtigen Leseraster kloniert waren, oder bei denen es sich um Teilsequenzen handelte, die nicht die von den Protein-spezifischen Antikörper erkannten Epitope umfaßten. Ungefähr 90% der Klone mit Inserts, die nicht im richtigen Leseraster kloniert waren, wurden durch den RGS·His Antikörper nicht erkannt.

Ein Protokoll für die Analyse von His_6-Fusionsproteinen durch *matrix assisted laser ionisation/desorption* Massenspektrometrie (MALDI-MS) wurde etabliert. His_6-Fusionsproteine wurden auf Ni^{2+}-Chelator-beschichteten magnetisierbaren Partikeln immobilisiert und so unter denaturierenden Bedingungen gereinigt und direkt mit Trypsin verdaut. Die Massen der tryptischen Peptide wurden gemessen und mit vorhergesagten Massen aus Sequenzdatenbanken verglichen.

Die Verwendung von geordneten cDNA Expressionsbanken erweitert die Anwendung von geordneten DNA-Banken auf Protein-basierende Screening-Methoden, und schafft so eine direkte Verbindung zwischen experimentellen Daten, die auf DNA- und Proteinbasis

gewonnen werden. Die hEx1 Expressionbank ermöglicht eine schelle Identifikation von Expressionsklonen für bestimmte Gene durch Screening mit DNA-Hybridisierungs-Sonden oder Antikörpern. Die Protein-Produkte von Klonen der cDNA-Bank können direkt für funktionelle Assays verwendet werden, oder können, falls *inclusion bodies* gebildet werden, als Antigene für die Produktion von Antikörpern verwendet oder möglicherweise in vitro gefaltet werden. Bibliotheken aus Expressionsklonen könnten, in Kombination mit Protein-Expression und -Reinigung in Mikrotiterplatten mit hohem Durchsatz, einen sinnvollen Weg zur Herstellung von Gen-Produkt-Katalogen darstellen.

8. Appendix

8.1 Abbreviations

2D-PAGE	two-dimensional polyacrylamide gel electrophoresis
A	adenine
A_{280}	absorbance at 280 nm
aa	amino acid
AMP	adenosine-5'-monophosphate
AP	alkaline phosphatase
APS	ammonium persulfate
ATP	adenosine-5'-triphosphate
BCIP	5-bromo-4-chloro-3-indolyl-phosphate
BMP	bone morphogenetic protein
bp	base pairs
BSA	bovine serum albumin
c	concentration
C	cytosine
CaM	calmodulin
CE	capillary electrophoresis
cAMP	3'-5'-cyclic AMP
cDNA	complementary DNA
Ci	Curie
COX4	cytochrome-c oxidase subunit IV
cpm	counts per minute
d	Dalton
dATP	deoxyadenosine-5'-triphosphate
DEPC	diethylpyrocarbonat
DIG	digoxigenin
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dNTP	deoxyribonucleosid-5'-triphosphate
dpm	radioactive decays per minute
dsDNA	double stranded DNA
DTT	1,4-dithiothreitol
ε	absorption coefficient
E. coli	*Escherichia coli*
EDTA	ethylenediaminetetraacetic acid
EGTA	ethyleneglycol-bis-(2-aminoethylether)-N,N,N',N'-tetraacetic acid
ELISA	enzyme-linked immunoabsorbent assay
g	acceleration due to gravity, g = 9.81 ms^{-2}
G	guanine
β-Gal	β-galactosidase, hydrolysis of lactose to glucose and galactose
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
HEPES	N-(2-hydroxyethyl)piperazine-N'-2-(ethanesulfonic acid)
hEx1	human expression library 1
hIk-1	human ikaros 1 gene
His	L-Histidine
HSP	heat shock protein
IgG	immunoglobulin G
IPTG	isopropyl-β-D-thiogalactopyranosid
IMAGE	integrated molecular analysis of genomes and their expression
kbp	kilo base pairs
kd	kilo Dalton
lacI	*lac* repressor gene

lacI^{Q}	*lacI* promoter mutation leading to enhanced expression
lacZ	β-galactosidase gene
LDH	lactate dehydrogenase
M	mole/litre
MALDI	matrix assisted laser desorption/ionisation
MALDI-MS	MALDI mass spectrometry
mKd1	mouse kidney library 1
MOPS	3-(N-morpholino)propanesulfonic acid
MPI-MG	Max-Planck-Institut für Molekulare Genetik, Berlin
mRNA	messenger RNA
MS	mass spectrometry
n.d.	not determined
NAD^+	β-nicotinamide adenine dinucleotide, oxidised
NADH	β-nicotinamide adenine dinucleotide, reduced
NBT	4-nitro blue tetrazolium chloride
NMR	nuclear magnetic resonance
NTA	nitrilotriacetic acid
OD_{600}	optical density at 600 nm
oligo(dT)	oligo-deoxythymidine
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
p	phosphate
PCR	polymerase chain reaction
PDE	phosphodiesterase
PE-C	pyridylethyl-cysteine
PEG	polyethyleneglycol
PK	pyruvate kinase
poly(A)	polyadenylic acid
PVDF	polyvinylidene difluoride
RBS	ribosome binding site
RGS	arginine-glycine-serine
RNA	ribonucleic acid
RNase	ribonuclease
rpm	revolutions per minute
RT	reverse transcriptase
RXR	retinoic acid X receptor
RZPD	Ressourcen-Zentrum Primär-Datenbank
SAP	shrimp alkaline phosphatase
SA-PMP	streptavidin paramagnetic particles
SDS	sodium dodecylsulfate
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
SSC	saline sodium citrate
ssDNA	single stranded DNA
T	thymine
TAE	Tris-acetate-EDTA
Taq	DNA polymerase from *Thermus aquaticus*
TBE	Tris-borate-EDTA
TBS	Tris-buffered saline
TCA	trichloroacetic acid
TCEP-HCl	tris(2-carboxyethyl)-phosphine hydrochloride
TE	Tris-EDTA
TEA	triethanolamine
TEMED	N,N,N',N'-tetramethylethylendiamine
TEN	Tris-EDTA-NaCl
TFA	trifluoroacetic acid
TOF	time of flight
Tris	tris(hydroxymethyl)aminomethan
tRNA	transfer RNA
U	units

URL	uniform resource locator
V	volume
VDAC1	voltage-dependent anion channel isoform 1
v/v	volume per volume
w/v	weight per volume
YAC	yeast artificial chromosome

8.2 Curriculum vitae

Name Konrad Büssow

E-mail buessow@mpimg-berlin-dahlem.mpg.de

Ausbildung

Sept. 1981 – Juni 1988	Schadow Schule Berlin, ABITUR
Okt. 1988 – März 1990	Technische Universität Berlin, Studiengang Biotechnologie
April 1990 – April 1995	Freie Universität Berlin, Studiengang Biochemie, DIPLOM BIOCHEMIE
13. Okt. 1994 – 13. April 1995	Diplomarbeit „Klonierung und Sequenzierung der Promotoren der DNA-Reparaturgene hMLH1 und hMSH2, Suche nach Polymorphismen“ am Welcome Trust Centre for Human Genetics, Oxford
Mai 1995 – heute	Doktorand am Max-Planck-Institut für Molekulare Genetik Berlin, Abteilung Prof. Dr. H. Lehrach, Arbeitsgruppe Dr. G. Walter

8.3 Publications

Parts of this work have been published (129,130). A patent has been filed describing techniques established in this work (131).