Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Parallel cloning, expression, purification and crystallization of human proteins for structural genomics

Hai-tao Ding,[a] Hui Ren,[a] Qiang Chen,[a] Gang Fang,[a] Lan-fen Li,[a] Rui Li,[a] Zhuo Wang,[a] Xiao-yu Jia,[a] Yu-he Liang,[a] Mei-hao Hu,[a] Yi Li,[a] Jing-chu Luo,[a] Xiao-cheng Gu,[a] Xiao-dong Su,[a,b] Ming Luo[a,c] and Shan-yun Lu[a]*

[a]Laboratory of Structural Biology, College of Life Sciences, Peking University, Beijing 100871, People's Republic of China, [b]Department of Molecular Biophysics, KC, PO Box 124 Lund University, SE-221 00 Lund, Sweden, and [c]Department of Microbiology, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA

Correspondence e-mail: lusy@lsc.pku.edu.cn

Received 28 January 2002
Accepted 10 September 2002

54 human genes were selected as test targets for parallel cloning, expression, purification and crystallization. Proteins from these genes were selected to have a molecular weight of between 14 and 50 kDa, not to have a high percentage of hydrophobic residues (*i.e.* more likely to be soluble) and to have no known crystal structures and were not known to be subunits of heterocomplexes. Four proteins containing transmembrane regions were selected for comparative tests. To date, 44 expression clones have been constructed with the Gateway™ cloning system (Invitrogen, The Netherlands). Of these, 35 clones were expressed as recombinant proteins in *Escherichia coli* strain BL21 (DE3)-pLysS, of which 12 were soluble and four have been purified to homogeneity. Crystallization conditions were screened for the purified proteins in 96-well plates under oil. After further refinement with the same device or by the hanging-drop method, crystals were grown, with needle, plate and prism shapes. A 2.12 Å data set was collected for protein NCC27. The results provide insights into the high-throughput target selection, cloning, expression and crystallization of human genomic proteins.

1. Introduction

Following the completion of a large number of genome sequences and in anticipation of the completion of the human genome sequence (Lander *et al.*, 2001; Venter *et al.*, 2001; Glaser *et al.*, 2001), much attention is now shifting towards the functions of the gene products (Fields *et al.*, 1999; Durick *et al.*, 1999). These functions can be inferred by various approaches, such as gene trapping, gene knockout or knock-in, transgenics or yeast two-hybrid systems (Galli-Taliadoros *et al.*, 1995; Bai & Elledge, 1996). Since the functions of proteins are correlated with their three-dimensional folds, structure determination can be incorporated into the process of function assignment (Shapiro & Harris, 2000; Teichmann *et al.*, 2001; Moult & Melamud, 2000; Skolnick *et al.*, 2000; Burley *et al.*, 1999). However, there is presently a large disparity between the number of proteins in the human genome (estimated at about 35 000; Lander *et al.*, 2001; Venter *et al.*, 2001) and the number of human proteins for which three-dimensional structural information is available (3041 *Homo sapiens* structures in the PDB on 7 December 2001; Berman *et al.*, 2000). The PDB contains only ~600 (SCOP; Murzin et al., 1995) different folds out of an estimated 1000–10 000 (Brenner *et al.*, 1997; Wolf *et al.*, 2000). Since it is rare to find novel folds among newly determined structures by traditional routines, it is necessary to determine new protein structures by high-throughput methods (Ryu & Nam, 2000).

Several factors have made it possible to initiate structural gemomics research in a high-throughput manner: the recent advent of bioinformatics (Luscombe *et al.*, 2001), the application of a variety of expression systems including *E. coli* (Baneyx, 1999), yeast (Buckholz, 1993) or baculovirus/insect (Griffiths & Page, 1997) and the development of new methods in X-ray crystallography (Beauchamp & Isaacs, 1999) such as cryofreezing of protein crystals, robotic crystallization screening, selenomethionine derivatization, single- or multiwavelength anomalous methods (Ealick, 2000), automated data processing (Abola *et al.*, 2000) and improvements in structure-prediction programs. The ultimate goal of structural genomics is to provide structural information on all known proteins (Mittl & Grutter, 2001). The improved completeness of protein structure data will help in the prediction of protein structures and the discovery of functional relationships of proteins to their structures (Teichmann *et al.*, 2001; Kim, 1998), as well as protein engineering of industrial enzymes and structure-based drug design (Gane & Dean, 2000; Klebe, 2000).

Before it is practical to determine protein structures on a large scale (Service, 2000), it is necessary to develop technological platforms for high-throughput target selection (Vitkup *et al.*, 2001; Sali, 2001), gene cloning, protein expression, protein purification, crystallization and X-ray or nuclear magnetic resonance structure determination. Pilot structural genomics projects have been started with targets mostly selected from prokaryotic genes in order to look for new protein folds (Christendat *et al.*, 2000; Burley *et al.*, 1999) or proteins related to key pathways or diseases (Terwilliger, 2000). In this study, we have performed a pilot project on parallel methods to produce and crystallize human proteins. Our results demonstrate that the parallel scheme that we have assembled based on currently available techniques, including Gateway™ cloning technology (Walhout *et al.*, 2000) for gene cloning and protein expression and a 96-well plate under-oil method for crystallization screening (http://www-structure.llnl.gov/crystool), is feasible for our large-scale project.

2. Materials and methods

2.1. Materials

cDNA clones for the preparation of entry vectors were obtained from Drs Zhu Chen and Ze-Guang Han (Center for the Human Genome Program in Southern China) and the IMAGE library (Lennon *et al.*, 1996). Gateway™ BP Clonase Enzyme Mix, Gateway™ LR Clonase Enzyme Mix, donor vector pDONR201, *E. coli* strains of library efficiency DH5α competent cells (Tang *et al.*, 1994) and BL21 (DE3)-pLysS competent cells (Derman *et al.*, 1993) were products from Invitrogen (The Netherlands). Destination vector pET11a-DEST was constructed by Dr Xin-li Lin (Oklahoma Medical Research Foundation) from pET11a for its higher yield of protein expression (Studier *et al.*, 1990). *Pfu* DNA Polymerase and Wizard Plus SV Minipreps DNA purification systems were from Promega (Madison, USA). Isopropylthio-β-D-galactoside (IPTG), ampicillin and kanamycin were from Sangon (Shanghai, People's Republic of China). ÄKTA FPLC and Sephacryl-100, Hitrap-Q/S, Resource-Q/S and Superdex-75 columns were from Amersham Pharmacia Biotech Asia Pacific Ltd (Hong Kong, People's Republic of China). 96-well Nuclon Surface plates (Alga Nunc International, Denmark) were used for crystallization screening under oil.

2.2. Methods

2.2.1. Bioinformatic analysis and primer design. 54 human genes were selected by several criteria. Firstly, the molecular weights (MW) of the proteins encoded using these genes were calculated and those with MWs between 14 and 50 kDa were accepted. Secondly, genes with full-length cDNA clones that were available in our human cDNA libraries, whose proteins do not have high percentage of hydrophobic residues and are not known to be subunits of heterocomplexes were retained. Thirdly, genes for which no crystal structures were available for their proteins were retained. In addition, six genes related to leukaemia were selected by applying more flexible criteria.

A pair of primers that correspond to the first and the last 24 nucleotides of the gene and additional nucleotides corresponding to the sequences required by the Gateway™ BP clonase reaction (Gateway Cloning Technology, Invitrogen, The Netherlands) were designed. Primer 1: GGGGACAAGTTTGTACAAAAAAGCAGGCTTA + 24 bp gene-specific nucleotides. Primer 2: GGGGACCACTTTGTACAAGAAAGCTGGGTC + 24 bp gene-specific nucleotides (complement strand). The transmembrane domain and signal peptide regions in the genes were predicted with the computer programs *TMHMM* (Krogh *et al.*, 2001; Sonnhammer *et al.*, 1998) and *SignalP* (Nielsen *et al.*, 1997, 1999). Predicted membrane proteins were either kept as intact molecules or truncated to produce an ectomembrane domain. Signal peptides were deleted from the proteins through primer design.

2.2.2. Gene cloning. DNA fragments of these genes were amplified by the polymerase chain reaction (PCR) at 367 K for 5 min; 367 K for 1 min, 323 K for 1 min and 345 K for 2 min for 30 cycles, then 345 K for 10 min. The fidelity of DNA synthesis was enhanced using *Pfu* DNA polymerase. After the reaction, the PCR products were purified as follows. (*a*) Add 150 µl TE (10 m*M* Tris–HCl, 1 m*M* EDTA pH 8.0) per 50 µl PCR mixture. (*b*) Mix well and then add 100 µl 30% polyethylene glycol (PEG) 8000/30 m*M* $MgCl_2$. (*c*) Mix immediately and centrifuge at 12 000 rev min^{-1} for 15 min at 277 K. (*d*) Remove the supernatant and dissolve the pellet in 50 µl TE and then check for recovery on agarose-gel electrophoresis.

Purified DNA fragments of the genes were then inserted into a donor vector pDONR201 by BP recombination reaction (*i.e.* a recombination reaction through *att*B and *att*P recombination sites) to create entry clones. The 10 µl BP cloning reaction system (including 2 µl BP reaction buffer, 1 µl pDONR201 vector at 150 ng µl^{-1}, 2 µl PCR product at 20 ng µl^{-1}, 2 µl BP Clonase Enzyme Mix, TE added to 10 µl) was incubated at 298 K for 2 h. 2 µl of BP reaction was transformed into 100 µl of library efficiency DH5α competent cells. The mixture was placed on ice for 30 min. The cells were subjected to heat shock at 315 K for 50 s, followed by sitting on ice for 1–2 min. The sample was then diluted with 450 µl Luria–Bertani (LB) medium (10 g tryptone, 5 g yeast, 10 g NaCl in 1 l distilled water, 1.5 g agar to give a solid medium) and incubated at 310 K for 1 h. Transformants were selected on LB plates containing 50 µg ml^{-1} kanamycin. The positive clones were confirmed by colony PCR. Finally, target genes in entry clones were transferred into the destination vector pET11a-DEST *via* an LR recombination reaction (a recombination reaction through *att*L and *att*R recombination sites) to create expression clones as the follows. 10 µl LR cloning reaction system (including 2 µl LR reaction buffer, 1 µl linearized pET11a-DEST vector at 150 ng µl^{-1}, 2 µl entry clone at 50 ng µl^{-1}, 2 µl LR Clonase Enzyme Mix, TE added to 10 µl) was incubated at 298 K for 2 h. LR products were then transformed into library efficiency BL21 (DE3)-pLysS competent cells as above. Transformants were selected in LB plates containing 100 µg ml^{-1} ampicillin. The positive clones were confirmed by colony PCR and used for protein expression.

2.2.3. Protein expression. Recombinant proteins were expressed in *E. coli* strain BL21 (DE3)-pLysS as follows. The isolated expression clones were inoculated into 5 ml ZB (10 g NZ-amine, 5 g NaCl in 1 l distilled water; Sambrook *et al.*, 1989) liquid medium containing 100 µg ml^{-1} of ampicillin. They were then cultured on a roller/incubator at 200 rev min^{-1} at 303 K overnight. 100 µl overnight bacterial cultures were added to 5 ml LB liquid medium containing 100 µg ml^{-1} ampicillin. They were cultured at 200 rev min^{-1} at 310 K until the OD$_{600}$ value of the bacterial cultures reached nearly 0.8. IPTG was added to a final concentration of 0.5 m*M* and continuously cultured in the same way. Control samples were processed under the same conditions without adding IPTG. Individual 1 ml aliquots were taken 5 and 12 h after induction and were centrifuged at 12 000 rev min^{-1} for 30 s. The pellets were suspended in 100 µl 1× SDS loading buffer [50 m*M* Tris–HCl pH 6.8, 100 m*M* dithiothreitol (DTT), 2% SDS, 0.1% bromophenol blue, 10% glycerol] and boiled for 5 min before being loaded onto SDS–PAGE. The expression results were evaluated by comparing those samples with controls to observe whether additional bands corresponding to target genes were present.

2.2.4. Soluble protein or inclusion-body identification. 200 ml of each bacterial culture containing a recombinant protein was centrifuged at 5000 rev min^{-1} for 10 min. The pellet was suspended in 15 ml TN buffer (50 m*M* Tris, 150 m*M* NaCl pH 7.5) and sonicated (on for 3 s, off for 3 s; 100 cycles), 1 ml aliquot of lysate was taken and centrifuged at 12 000 rev min^{-1} for 30 min. The pellet was dissolved in 8 *M* urea (8 *M* urea, 0.1 *M* Tris, 1 m*M* glycine, 1 m*M* EDTA, pH 10). The recombinant protein in the supernatant or pellet was analyzed by SDS–PAGE.

2.2.5. Protein purification. 2 l of bacterial culture was centrifuged at 5000 rev min^{-1} for 10 min. The pellet was suspended in 20 ml buffer system selected according to the start buffer to be used in the ion-exchange chromatography purification of the next step and sonicated as above. The lysate was clarified by centrifuging at 10 000 rev min^{-1} for 1.5 h and the supernatant was loaded onto a gel-filtration Sephacryl-100 column. Different peak fractions were collected with a flow rate of 40 ml h^{-1} for one and half column volumes and checked by SDS–PAGE. Those containing the recombinant protein were pooled together. Pooled fractions were then loaded onto a FPLC ion-exchange Hitrap-Q/S or Resource-Q/S column according to their calculated isoelectric point (pI). The pH of the running buffer was set to be more than one unit away from the pI of the recombinant protein. The recombinant protein was eluted from the column with a linear gradient of 0–0.5 *M* NaCl or a refined gradient that gave better separation for the target protein. Fractions were evaluated by SDS–PAGE and the peak containing the recombinant protein was collected. When necessary, the purity of recombinant proteins was further improved by running through an FPLC Superdex-75 column with a flow rate of 0.5 ml min^{-1} for 1.5 column volumes. The concentration of the recombinant proteins was finally measured with a Bio-Rad Protein Assay kit (Bio-Rad Pacific Ltd).

2.2.6. Crystallization and data collection. Crystallization conditions were screened for each of the purified proteins in 96-well plates under oil. After mixing 2 µl protein solution with 2 µl of a precipitant solution (one of 192 different conditions), the drop was covered with 200 µl of a mineral/silicon oil mixture. 96 random conditions and 96 gradient conditions were created with *Crystool* (Segelke, 1995; http://www-structure.llnl.gov/crystool) according to selected buffer systems, precipitants, salt concentrations and additives. The 96 random conditions for protein crystallization are designed with the following parameters.

(i) Precipitants: PEG 3350, ammonium sulfate, 2-methyl-2,4-pentanediol (MPD), sodium chloride.

(ii) Buffer: 2-(*N*-morpholino)ethanesulfonic acid (MES) pH 6.5, *N*-(2-hydroxyethyl)piperazine-*N*′-(2-ethanesulfonic acid) (HEPES) pH 7.5, boric–borax pH 8.5, sodium citrate pH 5.0.

(iii) Additives: sodium nitrate, calcium chloride.

(iv) Detergent: *n*-octyl β-D-glucopyranoside (β-OG).

The 96 gradient conditions are designed with the following parameters.

(i) Precipitants: PEG 4000, ammonium sulfate, MPD, 2-propanol.

(ii) Buffer: sodium cacodylate pH 5.5, HEPES pH 7.5, 3-(cyclohexylamino)-1-propanesulfonic acid (CAPS) pH 9.0.

(iii) Additives: zinc chloride, sodium chloride, calcium chloride, sodium nitrate, DTT, magnesium acetate.

(iv) Detergent: β-OG

After small crystals were observed, further refinements of the conditions, either under oil or using the hanging-drop method, was carried out in order to grow crystals suitable for diffraction studies. Diffraction data were collected on a MAR Research image-plate system (Model Mar345) and an X-ray rotating-anode generator (Rigaku Model 2000, Japan).

3. Results and analysis

The 54 genes included in this work are listed in Table 1, with their names, GenBank accession numbers (Ac) and the presence of predicted transmembrane domains or signal peptides. The protein from Ac AAH07438 (RIKEN) was predicted to be a transmembrane protein with six transmembrane domains (Fig. 1). The protein from the 'defender against cell death gene' (Ac AAH07403) was predicted to be a transmembrane protein with three transmembrane domains. These genes were kept as single-chain proteins in the primer design. 8D6 antigen (Ac AAH07083) was predicted to be a membrane protein with a 200-residue inside fragment and two outside fragments with 30 and 29 residues. The 200-residue fragment was selected to be expressed. Reticulon 4 (Ac AAH07109) was predicted as a membrane protein with an outside fragment and two inside fragments. The outside fragment was suggested to be the active component (GrandPre *et al.*, 2000) and it was selected as the protein to be expressed (Fig. 1). Four membrane proteins were included here for comparative tests to see whether they were suitable for expression in this system. Neuromedin B (Ac AAH07431), collagenase inhibitor (Ac AAH07097) and sememogelin (Ac AAH07096) were predicted to have 26-, 23- and 23-residue signal peptides, respectively. Signal peptides of these proteins were deleted in recombinant protein expression through primer design. These proteins were also selected as controls. The results of gene cloning, expression and the number of proteins in soluble form are tabulated in Table 2 (excluding the four transmembrane proteins). This showed that 70% of the targeted genes could be expressed as recombinant proteins. The success rate of this expression system is quite suitable for structural genomics projects.

Table 1
List of 54 target genes.

Ac, accession number in GenBank; TM, transmembrane protein predicted by the program *TMHMM*; SP, protein with signal peptide predicted by the program *SignalP*.

No.	Name	Ac	TM or SP
1	Interleukin 18	BC007461	
2	Acid phosphatase 1	AAH07422	
3	Unknown	BC007459	
4	Hypothetical	BC007457	
5	GDP-mannose pyrophosphorylase A	BC007456	
6	IκB-interacting Ras-like protein 2	AAH07450	
7	Hypothetical	AAH07446	
8	Neighbour of COX4	AAH07445	
9	RIKEN	AAH07438	TM
10	Interferon-related developmental regulator 2	AAH07437	
11	Unknown	AAH07434	
12	Neuromedin B	AAH07431	SP
13	Arrestin	AAH07427	
14	Adrenal gland protein	AAH07426	
15	Unknown	AAH07423	
16	Two-pore channel	AAH07419	
17	RIKEN	AAH07416	
18	UBX domain	AAH07414	
19	Unknown	BC007410	
20	Defender against cell death	AAH07403	TM
21	Unknown	AAH07394	
22	Dactylaplasia	AAH07380	
23	HSCARG	AAH07364	
24	Dicarboxylate transporter	AAH07355	
25	Uridine phosphorylase	AAH07348	
26	X breakpoint 2	AAH07343	
27	Short-chain alcohol dehydrogenase	AAH07339	
28	BM039	AAH07334	
29	Macrophin	AAH07330	
30	Pyrophosphate phosphatase	AAH07324	
31	Sortilin	AAH07296	
32	Nuclear-distribution gene	AAH07280	
33	Px19	AAH07268	
34	Myristoyltransferase	AAH07258	
35	Pelota	AAH07249	
36	Ras homolog gene	AAH07245	
37	Mesenchymal stem-cell protein DSC92	AAH07222	
38	Rad50-interacting protein	AAH07120	
39	DNA-fragmentation factor	AAH07112	
40	Reticulon 4	AAH07109	TM
41	Translocase	AAH07106	
42	Small acidic protein	AAH07103	
43	Rcd1	AAH07102	
44	Collagenase inhibitor	AAH07097	SP
45	Semenogelin	AAH07096	SP
46	8D6 antigen	AAH07083	TM
47	Dolichyl-phosphate mannosyltransferase	AAH07073	
48	Lsm3 protein	AAH07055	
49	NADH ubiquinone reductase 24 kDa subunit	M22538	
50	Thiopurine methyltransferase	S62904	
51	Human COP9 homologue	U51205	
52	Human 14-3-3 epsilon	U54778	
53	Nuclear chloride ion-channel protein	U93205	
54	Ribosomal protein L7	X52967	

47 of the target genes could be amplified by PCR under standard conditions. The two failures might be caused by those pairs of primers tending to form dimers more easily than complementing gene templates, or it may be that the designed primers did not match the gene sequences perfectly (Brownie *et al.*, 1997). These seven failed genes were tested with different annealing temperatures between 313 and 338 K; two could be amplified with new conditions. The 49 final PCR products were then used in a BP reaction to construct entry clones. The BP reaction is the first key step for gene cloning because entry clones are more difficult to acquire than expression clones. The success of the BP reaction is determined mostly by the purity of PCR products and the amount of DNA used in the BP reaction system. By carefully adjusting these parameters, the 44 entry clones were successfully

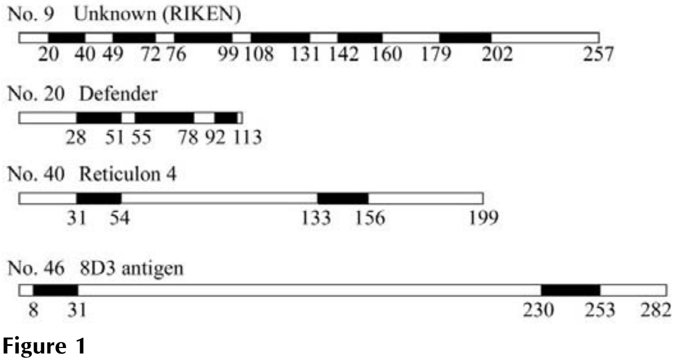

Figure 1
Transmembrane regions of proteins Nos. 9, 20, 40, 46 predicted by *TMHMM*. Protein sequences are shown by bars, with transmembrane peptides within them filled in black. Residue positions are indicated below the bars.

Table 2
Statistics of gene cloning, expression and percentage of soluble proteins from 50 genes.

Four transmembrane proteins were excluded from the table.

	Target genes	Expression clones	Expressed proteins	Soluble proteins
Quantity	50	44	35	12
Percentage of previous step (%)		88.0	79.5	34.3
Percentage of all genes (%)		88.0	70.0	24.0

obtained. Using entry clones in the LR reaction, each yielded an expression clone. The 44 expression clones were tested for protein expression and 35 of the 44 showed specific bands on SDS–PAGE gels. 12 soluble or partially soluble proteins were observed from these 35 expressed recombinant proteins (Fig. 2) and four of them were purified to homogeneity (Fig. 3).

Protein characteristics were predicted with Lasergene software (DNASTAR Inc.) and are summarized in Table 3. Average MW, pI and frequency of certain types of residues such as charged (Arg, Lys, His, Tyr, Cys, Asp, Glu), acidic (Asp, Glu), basic (Lys, Arg), polar (Asn, Cys, Gln, Ser, Thr, Tyr) and hydrophobic (Ala, Ile, Leu, Phe, Trp, Val) residues are listed in the table. The data show that the MW is closely correlated with protein expression in our expression system. Proteins with lower MW seem to be expressed better than those with higher MW. All 14 molecules with MW below 25 kDa, except neuromedin B, were expressed (92.9% success). For 28 molecules with MW between 25 and 40 kDa, 19 were expressed (67.9% success). On the other hand, five of the eight proteins with MW higher than 40 kDa could not be expressed (37.5% success).

There are various reasons for failure or poor expression of eukaryotic genes in an *E. coli* expression system. Possible reasons include that (i) the sequence of a gene at the 5′ terminus is not suitable for initiating translation in the T7 RNA polymerase/promoter expression system, (ii) codons in the sequence of a gene are rare codons of *E. coli*, (iii) expressed recombinant protein is degraded by proteinases in *E. coli* or (iv) the translation is terminated before completion. Since strain BL21 (DE3)-pLysS is deficient in both lon and ompT proteases (Derman *et al.*, 1993) and the N-terminal fusion of target genes in pET11a-DEST favors the T7 RNA polymerase/promoter expression system, the main reasons for the failure of this system to express target genes may be codon usage differences between human genomic genes and the *E. coli* translation mechanism.

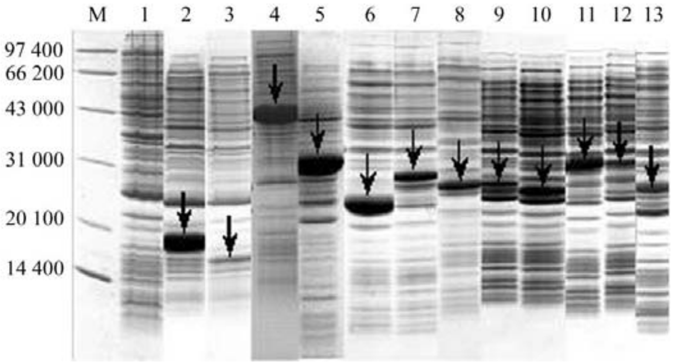

Figure 2
35 recombinant proteins were tested for their expression form, *i.e.* as either soluble proteins or inclusion bodies. Four of the 35 recombinant proteins were expressed in soluble form at 310 K, while eight of the remaining 31 were expressed as soluble proteins at 291 K. Lane M, protein molecular-weight markers (Da); lane 1, total bacterial proteins as controls without IPTG induction; lanes 2–13, supernatants of *E. coli* lysate containing recombinant proteins, which in turn refer to Nos. 2, 21, 39, 53 at 310 K and 1, 3, 8, 11, 14, 23, 25, 27 at 291 K.

Table 3
Relationship between expression of 50 target proteins by parallel method and calculated characteristics of these proteins.

Average MW, pI and percentage of certain type of residues are listed for not expressed or expressed proteins, proteins expressed as inclusion bodies or soluble ones. (Four transmembrane proteins were excluded from the table.)

	MW	pI	Charged residues (%)	Acidic residues (%)	Basic residues (%)	Polar residues (%)	Hydro-phobic residues (%)
Not expressed (15)	35677	8.47	31.04	10.66	13.86	25.01	31.99
Expressed (35)	27845	7.53	31.92	12.24	12.73	26.64	30.53
Inclusion body (23)	27481	7.74	31.62	11.98	12.84	27.06	29.57
Soluble (12)	28543	7.13	32.5	12.74	12.52	25.83	32.37

Heterogeneous recombinant proteins expressed in large amounts in the host *E. coli* strain could easily cause the formation of insoluble inclusion bodies (Hoffmann *et al.*, 2001; Strandberg & Enfors, 1991) or poorly soluble products. In our protocol, three soluble proteins were identified from 35 expressed proteins that were induced at 310 K, while nine additional soluble proteins were found at 291 K. It is concluded that induction at lower temperature is a better choice for the expression of more soluble proteins from the Gateway$^{\text{TM}}$ system.

Crystallization screening under oil, using the commonly available 96-well plates, was conducted for the four purified proteins Nos. 1, 2, 23 and 53, using 96 randomly selected and 96 linear gradient conditions with 2 µl protein mixed with 2 µl

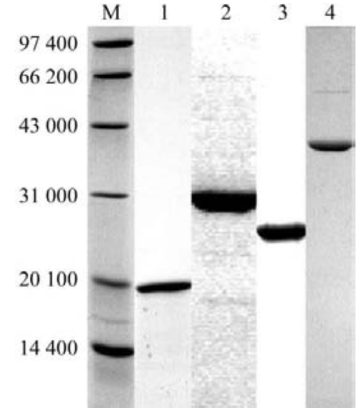

Figure 3
Four proteins were purified to homogeneity through Sephacryl-100, Hitrap-Q or S column and, if needed, an additional Superdex-75 column. Lane M, protein molecular-weight markers (Da); lanes 1–4, purified proteins Nos. 1, 2, 23, 53.

conditional buffer in each well. Each well was inspected weekly under an optical microscope. Crystals appeared within 3–5 d and grew to their final size in one week (Fig. 4). The loss of water in protein solutions under oil was faster than in the normal hanging-drop or sitting-drop methods. Since the screen under oil contained 192 different conditions and 2 µl protein solution for each well, 4.0 mg protein was sufficient for one complete screen. Conditions that yielded small crystals were further refined with finer steps around the initial conditions using a larger drop size (5 µl/5 µl) under oil or in the hanging-drop method. For instance, plate-shaped crystals were found for NCC27 (No. 53) in well No. 76 from 96 random conditions, which contained 0.1 *M* MES pH 6.5, 0.2 *M* $CaCl_2$, 0.79 *M* β-OG and 22.35% PEG 3350. Refinement included variations in the PEG concentration and the concentrations of $CaCl_2$ and β-OG. A large prismatic crystal grew to 0.2 × 0.2 × 0.8 mm under oil with a precipitation solution consisting of 0.1 *M* MES pH 6.5, 0.2 *M* $CaCl_2$, 0.30 *M* β-OG, 26% PEG. This crystal diffracted to 2.2 Å resolution when examined by X-ray diffraction and Table 4 lists its data-collection statistics. Needles were observed for proteins Nos. 1 and 2 from 96 random conditions and further refinement of these conditions is ongoing.

Table 4
Data-collection statistics of NCC27.

Values in parentheses are for the last resolution shell (2.28–2.20 Å).

Wavelength (Å)	1.542
Resolution (Å)	30.0–2.20 (2.28–2.20)
Completeness (%)	96.8 (94.7)
$R_{\rm merge}$† (%)	5.3 (31.9)
$I/\sigma(I)$	33 (4.6)
Space group	$P2_1$
Unit-cell parameters (Å)	a = 42.26, b = 69.78, c = 81.97, β = 90.08
No. of possible unique reflections	23404
No. of observed reflections	217399
$V_{\rm M}$ (Å^3 Da^{-1})	2.24
Solvent content for dimer (%)	45

† $R_{\rm merge} = \sum |I_{\rm obs} - I_{\rm avg}| / \sum I_{\rm obs}$, where the summation is over all reflections.

4. Discussion and concluding remarks

In this report, a pilot structural genomics project characterized by parallel operation of target gene selection, gene cloning, protein expression, protein purification, crystallization and X-ray diffraction data collection has been carried out. The results demonstrate that the target-selection criteria and the Gateway™ cloning system are suitable for structural genomics research, as 64% of the selected genes could be expressed well in *E. coli*.

A MW of between 14 and 50 kDa is a reasonable size for protein structure determination by crystallography (Smith *et al.*, 1986). Most interestingly, proteins with MW < 25 kDa were expressed more frequently than those with MW > 25 kDa.

Among four genes for which the protein products were predicted to have transmembrane peptides, two genes were kept as a complete open reading frame to include the transmembrane domain and inside/outside domains. The other two genes were truncated to include only the longest inside or outside fragment. Expression clones were produced for the four genes, but no protein expression was observed. For three other genes whose proteins were predicted with signal peptides, the signal peptides were deleted through primer design. Two of these genes were expressed well with high yields. The results indicate that transmembrane proteins are difficult to express in our system but secreted proteins can be expressed well after deleting the signal peptide.

The Gateway™ cloning system was selected for its advantage of large-scale operation with the same conditions for any gene. At the same time, sub-cloned genes in entry clones can be easily transferred to different expression clones (Walhout *et al.*, 2000). Excluding the four membrane proteins, 70% (35/50) of the genes were expressed in the Gateway™ system. This shows that with careful selection of target genes, the *E. coli* expression system is suitable for the production of human genomic proteins with high efficiency. However, a large number of these genes were expressed as insoluble proteins (46%, 23/50) and only 24% (12/50) were soluble. Of the 12 soluble proteins, four proteins with high yields could be purified for crystallization screening. After comparison of methods for the purification of recombinant proteins, we suggest that, if possible, the Hitrap-S column is preferable for use, as most of the constitutive proteins encoded by the *E. coli* genome will be washed out with the starting buffer. When a Hitrap-Q column is used, proteins with lower pI are easier to purify because they seem to bind to the column tighter than most of the other constitutive proteins, which can be washed out at lower salt concentrations. This information could be incorporated in future target selection.

Under-oil screening for crystallization conditions can be easily carried out in common 96-well plates with manual multi-channel pipettes or a robotic system. A large screen consumes only a few milligrams of protein. Further refinement of crystal

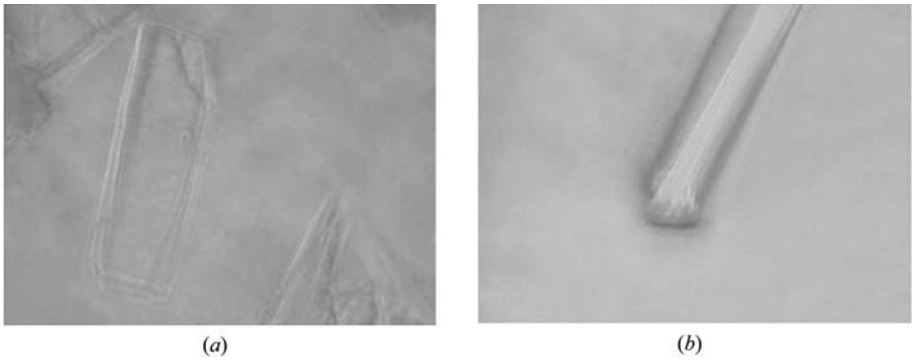

Figure 4
Photomicrographs of crystals of (*a*) protein No. 53 and (*b*) protein No. 2.

growth can be carried out under oil by using larger drop sizes. We found it difficult to translate the conditions for crystallization under oil to the hanging-drop vapour-diffusion method. Crystallization is a dynamic process, which can exhibit different characteristics in different setups.

We thank Drs Zhu Chen and Ze-guang Han for providing full-length cDNA clones, Drs James Finley and Song-lin Li for discussion of crystallization screens and Dr Zi-he Rao for use of the X-ray diffraction equipment in his laboratory. Last but not least, we also thank Dr Jian-li An for critical reading of the manuscript. The project is supported by the National High Technology (863) Program, the Ministry of Education and the 985 Program of Peking University.

References

Abola, E., Kuhn, P., Earnest, T. & Stevens, R. C. (2000). *Nature Struct. Biol.* **7** (*Suppl.*), 973–977.

Bai, C. & Elledge, S. J. (1996). *Methods Enzymol.* **273**, 331–347.

Baneyx, F. (1999). *Curr. Opin. Biotechnol.* **10**, 411–421.

Beauchamp, J. C. & Isaacs, N. W. (1999). *Curr. Opin. Chem. Biol.* **3**, 525–529.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Brenner, S. E., Chothia, C. & Hubbard, T. J. (1997). *Curr. Opin. Struct. Biol.* **7**, 369–376.

Brownie, J., Shawcross, S., Theaker, J., Whitcombe, D., Ferrie, R., Newton, C. & Little, S. (1997). *Nucleic Acids Res.* **25**, 3235–3241.

Buckholz, R. G. (1993). *Curr. Opin. Biotechnol.* **4**, 538–542.

Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999). *Nature Genet.* **23**, 151–157.

Christendat, D. *et al.* (2000). *Nature Struct. Biol.* **7**, 903–909.

Derman, A. I., Prinz, W. A., Belin, D. & Beckwith, J. (1993). *Science*, **262**, 1744–1747.

Durick, K., Mendlein, J. & Xanthopoulos, K. G. (1999). *Genome Res.* **9**, 1019–1025.

Ealick, S. E. (2000). *Curr. Opin. Chem. Biol.* **4**, 495–499.

Fields, S., Kohara, Y. & Lockhart, D. J. (1999). *Proc. Natl Acad. Sci. USA*, **96**, 8825–8826.

Galli-Taliadoros, L. A., Sedgwick, J. D., Wood, S. A. & Korner, H. (1995). *J. Immunol. Methods*, **181**, 1–15.

Gane, P. J. & Dean, P. M. (2000). *Curr. Opin. Struct. Biol.* **10**, 401–404.

Glaser, P. *et al.* (2001). *Science*, **294**, 849–852.

GrandPre, T., Nakamura, F., Vartanian, T. & Strittmatter, S. M. (2000). *Nature (London)*, **403**, 439–444.

Griffiths, C. M. & Page, M. J. (1997). *Methods Mol. Biol.* **75**, 427–440.

Hoffmann, F., Posten, C. & Rinas, U. (2001). *Biotechnol. Bioeng.* **72**, 315–322.

Kim, S.-H. (1998). *Nature Struct. Biol.* **5** (*Suppl.*), 643–645.

Klebe, G. (2000). *J. Mol. Med.* **78**, 269–281.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). *J. Mol. Biol.* **305**, 567–580.

Lander, E. S. *et al.* (2001). *Nature (London)*, **409**, 860–921.

Lennon, G., Auffray, C., Polymeropoulos, M. & Soares, M. B. (1996). *Genomics*, **33**, 151–152.

Luscombe, N. M., Greenbaum, D. & Gerstein, M. (2001). *Methods Inf. Med.* **40**, 346–358.

Mittl, P. R. & Grutter, M. G. (2001). *Curr. Opin. Chem. Biol.* **5**, 402–408.

Moult, J. & Melamud, E. (2000). *Curr. Opin. Struct. Biol.* **10**, 384–389.

Murzin, A. G., Brenner, S. E., Hubbard, T. J. P. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Nielsen, H., Brunak, S. & von Heijne, G. (1999). *Protein Eng.* **12**, 3–9.

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). *Protein Eng.* **10**, 1–6.

Ryu, D. D. & Nam, D. H. (2000). *Biotechnol. Prog.* **16**, 2–16.

Sali, A. (2001). *Nature Struct. Biol.* **8**, 482–484.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning. A Laboratory Manual*, 2nd ed. New York: Cold Spring Harbor Laboratory Press.

Segelke, B. (1995). PhD thesis, University of California, San Diego, USA.

Service, R. F. (2000). *Science*, **287**, 1954–1956.

Shapiro, L. & Harris, T. (2000). *Curr. Opin. Biotechnol.* **11**, 31–35.

Skolnick, J., Fetrow, J. S. & Kolinski, A. (2000). *Nature Biotechnol.* **18**, 283–287.

Smith, T. J., Kremer, M. J., Luo, M., Vriend, G., Arnold, E., Kamer, G., Rossmann, M. G., McKinlay, M. A., Diana, G. D. & Otto, M. J. (1986). *Science*, **233**, 1286–1293.

Sonnhammer, E. L., von Heijne, G. & Krogh, A. (1998). *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182.

Strandberg, L. & Enfors, S. O. (1991). *Appl. Environ. Microbiol.* **57**, 1669–1674.

Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990). *Methods Enzymol.* **185**, 60–89.

Tang, X., Nakata, Y., Li, H. O., Zhang, M., Gao, H., Fujita, A., Sakatsume, O., Ohta, T. & Yokoyama, K. (1994). *Nucleic Acids Res.* **22**, 2857–2858.

Teichmann, S. A., Murzin, A. G. & Chothia, C. (2001). *Curr. Opin. Struct. Biol.* **11**, 354–363.

Terwilliger, T. C. (2000). *Nature Struct. Biol.* **7** (*Suppl.*), 935–939.

Venter, J. C. *et al.* (2001). *Science*, **291**, 1304–1351.

Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001). *Nature Struct. Biol.* **8**, 559–566.

Walhout, A. J., Temple, G. F., Brasch, M. A., Hartley, J. L., Lorson, M. A., van den Heuvel, S. & Vidal, M. (2000). *Methods Enzymol.* **328**, 575–592.

Wolf, Y. I., Grishin, N. V. & Koonin, E. V. (2000). *J. Mol. Biol.* **299**, 897–905.