Protein production and purification

Structural Genomics Consortium[1–3], Architecture et Fonction des Macromolécules Biologiques[4], Berkeley Structural Genomics Center[5], China Structural Genomics Consortium[6,7], Integrated Center for Structure and Function Innovation[8], Israel Structural Proteomics Center[9], Joint Center for Structural Genomics[10,11], Midwest Center for Structural Genomics[12], New York Structural GenomiX Research Center for Structural Genomics[13–17], Northeast Structural Genomics Consortium[18,19], Oxford Protein Production Facility[20], Protein Sample Production Facility, Max Delbrück Center for Molecular Medicine[21], RIKEN Structural Genomics/ Proteomics Initiative[22] & SPINE2-Complexes[23,25]

In selecting a method to produce a recombinant protein, a researcher is faced with a bewildering array of choices as to where to start. To facilitate decision-making, we describe a consensus 'what to try first' strategy based on our collective analysis of the expression and purification of over 10,000 different proteins. This review presents methods that could be applied at the outset of any project, a prioritized list of alternate strategies and a list of pitfalls that trip many new investigators.

Recombinant proteins are used throughout biological and biomedical science. Their production was once the domain of experts, but the development of simple, commercially available systems has made the technology more widespread. As a result, also more widespread is an appreciation of the difficult, strategic choices inherent to the process. Commonly confronted questions include: should the protein(s) be expressed in bacteria, in yeast, in insect cells or in human cells? Which expression vector should be used? If bacterial expression is used, which strain(s) should be chosen? Should one express the full-length protein or a fragment thereof? Should the protein be tagged, and which affinity tag is the best? What is a good purification strategy, and what are the common pitfalls? Unfortunately, because every protein is different, there can be no 'right' answer to any of these questions *a priori*, and purification protocols and strategies must be worked out for each individual protein and with an eye to its intended use. This said, each project must begin somewhere, and purification strategies can now be

[1]Karolinska Institutet, Schéeles väg 2, 171 77 Stockholm, Sweden. [2]University of Oxford, Old Road Campus, Roosevelt Drive, Headington, Oxford OX3 7DQ, UK. [3]University of Toronto, 100 College St., Toronto, Ontario M5G 1L6, Canada. [4]Architecture et Fonction des Macromolécules Biologiques, Centre National de la Recherche Scientifique, Case 932, 163 Avenue de Luminy, 13288 Marseille Cedex 09, France. [5]Lawrence Berkeley National Laboratory and Department of Chemistry, University of California, 351A Donner Laboratory, Berkeley, California 94720, USA. [6]Tsinghua University, Beijing 100084, China. [7]University of Science and Technology of China, Hefei 230027, China. [8]Los Alamos National Laboratory, Mailstop M888, Los Alamos, New Mexico 87507, USA. [9]Weizmann Institute of Science, 2 Herzel Street, Rehovot 76100, Israel. [10]The Scripps Research Institute, 10550 N. Torrey Pines Rd., La Jolla, California 92037, USA. [11]Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California, 92121, USA. [12]Biosciences Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439, USA. [13]SGX Pharmaceuticals, Inc., 10505 Roselle Street, San Diego, California 92121, USA. [14]Department of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA. [15]Biology Department, 463, Brookhaven National Laboratory, Upton, New York 11973, USA. [16]Case Western Reserve University, 10900 Euclid Ave., Cleveland, Ohio 44016, USA. [17]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, University of California, San Francisco, 600 16th Street, San Francisco, California 94143, USA. [18]Center for Advanced Biotechnology and Medicine, Rutgers University, 679 Hoes Lane, Piscataway, New Jersey 08854, USA. [19]Department of Biological Sciences, Columbia University, 701 Fairchild Building, MC 2451, New York, New York 10027, USA. [20]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX37BN, UK. [21]Max Delbrück Center for Molecular Medicine (MDC), Robert-Rössle-Str. 10, 13092 Berlin, Germany, [22]Protein Research Group, Genomic Sciences Center, Yokohama Institute, RIKEN, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan. [23]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX37BN, UK. [24]Present addresses: Structural Biology, Helmoltz Center for Infection Research, Inhoffenstr. 7, 38124 Braunschweig, Germany (K.B.), and Department of Biology, and Center for Genomics and Systems Biology, New York University, 1009 Sliver Center, 100 Washington Square East, New York, NY 10003, USA (K.C.G.). [25]A complete list of authors appears at the end of this paper. Correspondence should be addressed to A.E. (aled.edwards@utoronto.ca).

PUBLISHED ONLINE 30 JANUARY 2008; DOI:10.1038/NMETH.F.202

Table 1 | Overview of targeted proteins

Organism	Targets cloned	Targets purified	Percentage purified
Viruses	335	118	35
Archaea	8,043	2,917	36
Bacteria	58,806	17,350	30
Eukarya	42,239	8,008	19

These data were obtained from TargetDB (21 December 2007) and include data from all structural genomics centers listed, plus updated information from the SGC. Source, http://targetdb.pdb.org/statistics/TargetStatistics.html and http://www.thesgc.com/structures/target_progress.php.

guided by evidence-based trends, probabilities and cautionary notes that have emerged from large-scale structural genomics studies. In this review, which is targeted to the researcher with limited experience in protein expression and purification, we draw on our collective experiences to suggest a 'consensus' starting point for soluble protein expression and purification.

Over the past decade, our laboratories have collectively targeted and purified tens of thousands of different proteins from the Eubacteria and Archaea, and thousands from the Eukarya, including fungal, nematode, parasite, plant and human proteins (**Table 1**). These proteins belong to many different classes, including proteins with no predictable structure, human proteins of therapeutic relevance, proteins from parasites and viruses, integral membrane proteins and multiprotein complexes. A near-complete list of these proteins is available in a database (TargetDB) maintained by the Protein Data Bank (PDB; http://targetdb.pdb.org/) under the auspices of the US National Institute of General Medical Sciences (NIGMS)-funded Protein Structure Initiative (http://www.nigms.nih.gov/Initiatives/PSI/). The European research network Structural Proteomics in Europe (SPINE) also provides detailed target lists online (http://www.spineurope.org/).

In efforts to identify an optimal approach(es) for the initial production and purification of a 'typical' protein, our groups have explored many different technologies and strategies. Our common objective has been to balance success rates with ease and breadth of use, speed, cost and versatility[1–16]. By comparing our independently optimized approaches, it is apparent that our preferred methods have, in many instances, evolved to be quite similar, but by no means identical (**Table 2**). Accordingly, in an effort to provide guidance to scientists interested in generating purified recombinant proteins, representatives from our research groups collaborated to articulate our 'consensus' advice (**Box 1**), along with a brief rationale for each choice. In essence, we tried to answer the question "what would you try first?", understanding that several choices are often possible or even desirable. We also provide guidance for those cases in which the initial attempt fails or problems are encountered, in other words, "What next?". In **Supplementary Methods** online, we provide links to online protocols offered by several structural genomics groups as well as detailed experimental protocols for the methods described here.

It is important to emphasize three aspects of this review. First, it is meant to serve as a guide to those members of the research community who are interested in expressing recombinant proteins, but who feel that they may not have the breadth of experience to decide among the various possible approaches. Second, we selected this consensus strategy because it is simple and has the widest use. There are other methods that are perhaps equivalent, but space limitations preclude an in-depth discussion of all possible cloning, expression and purification strategies. Third, the methods described here were developed with the intention to produce purified, soluble protein in close-to-milligram quantities; there are many applications for purified protein (biochemical assays, antibody production) that may not have such requirements.

Table 2 | Summary of approaches used by SG centers

Center	Main target sources	Cloning method	Expression promoter	Affinity tags	Small-scale expression method	Scale-up cultivation method	Purification strategy	Protein characterization	References
Structural Genomics Consortium	Human and human pathogens	LIC	T7	6His	96-well plates	1–2 l in Tunair shake flasks	IMAC and gel filtration using ÄKTA systems	ESI-MS	16,19
Architecture et Fonction des Macromolécules Biologiques	Mammalian viruses, higher eukaryotes, bacteria, phages	Gateway	T7	6His	96-well plates	1–5 l in shake flasks	IMAC and gel filtration using ÄKTA systems	MALDI-TOF	32,51
Berkeley Structural Genomics Center	Bacteria	LIC	T7	6His	96-well plates	1 l cultures in Fernbach flasks	IMAC and gel filtration using ÄKTA systems	DLS, MALDI, ANSEC	85
China Structural Genomics Consortium	Human	LIC, restriction enzyme–based	T7	6His	3 ml culture in test tubes	2 l in Erlenmeyer flasks	IMAC, and ion exchange chromatography and gel filtration	SDS-PAGE, DLS and mass spectrometry	8,9
Integrated Center for Structure and Function Innovation	*Mycobacterium tuberculosis, Bacillus subtilis, Thermotoga maritima*	Restriction enzymes and LIC	T7 and arabinose	6His	96-well plates	0.5–1 l in soda bottles or baffled shake flasks	IMAC and gel filtration	SDS-PAGE, densitometry, DLS and MALDI	

Table 2 | (continued)

Center	Main target sources	Cloning method	Expression promoter	Affinity tags	Small-scale expression method	Scale-up cultivation method	Purification strategy	Protein characterization	References
Israel Structural Proteomics Center	Higher eukaryotes, human pathogens	Restriction enzyme–based or LIC	T7	N-6His	4 ml culture in test tubes	0.5 l culture in 2-l shake flasks or 1.25 l culture in 5-l flasks (total 5–6 l per large-scale production)	IMAC, gel filtration, ion exchange chromatography and TEV cleavage	MALDI and ESI-MS	10
Joint Center for Structural Genomics	Bacteria	Polymerase incomplete primer extension	Arabinose	6His	96-well plates, ANSEC and mass spectromerty	Parallel 12–96 cultures in GNFermentor	IMAC, TEV cleavage, IMAC subtraction, ion exchange chromatography and gel filtration if necessary	ANSEC, LC-MS, SDS-PAGE	18,86–88
Midwest Center for Structural Genomics	Bacteria	LIC	T7	6His	96-well plates	Plastic bottles	IMAC, TEV cleavage, gel filtration	SDS-PAGE	89
New York Structural GenomiX Research Center for Structural Genomics	Human and >130 other species (ATCC and gene synthesis)	Topo (blunt)		C-6His and N-6His-Smt3	96-well plates	1–3 l shake flasks SeMet high yield	ÄKTAxpress and Ni-NTA column purification and gel filtration	MALDI and ESI-MS; protein identification by mass spectrometry and/or DNA sequencing	1
Northeast Structural Genomics Consortium	Prokaryotes and eukaryotes, including human	LIC	T7	6His	96-well plates	1–2 l in baffled shake flasks	IMAC, gel filtration using ÄKTAxpress, ion exchange chromatography if required	Caliper microfluidics, MALDI-TOF mass spectrometry, light scattering, NMR	13
Oxford Protein Production Group (SPINE)	Bacteria, human, viral pathogens	LIC	T7; β-actin or hCMV for mammalian cells	N- or C-6His	96-well plates; 25-cm^2 dishes for mammalian cells	1–2 l cultures	IMAC and gel filtration using ÄKTA systems	SDS-PAGE, ESI-MS, MALDI-TOF MS; LC-ESI-MS followed by ZIC-HILIC for glycosylated proteins	5,14,51,90
Protein Sample Production Facility, Max Delbrück Center for Molecular Medicine	Human and higher eukaryotes	Restriction enzyme–based, Gateway	T5 and T7	Mainly N-7His, occasionally N-GST or N-MBP	1–10 ml culture	1–8 l in shake flasks	IMAC and TEV cleavage, IMAC and gel filtration, and ion exchange chromatography using ÄKTA systems if necessary	Mass spectrometry, DLS	2,3,7
RIKEN Structural Genomics/ Proteomics Initiative	Human, mouse, bacteria, and archaea	Two-step PCR and TA cloning	T7	Histidine affinity tag (HAT)	30 μl in cell-free synthesis in 96-well plates	9–27 ml dialysis cell-free synthesis	IMAC and TEV cleavage, IMAC subtraction, ion exchange chromatography, gel filtration using ÄKTA systems if necessary	DLS, NMR, MALDI-TOF and quadrupole-TOF tandem mass spectrometry	72–76
SPINE2-Complexes	Human, viral proteins involved in subversion of human signaling pathways	LIC	T7; β-actin or hCMV for mammalian cells	N- or C-6His tag	96-well plates; 25-cm^2 dishes for mammalian cells	1–2 l cultures	IMAC and gel filtration using ÄKTA systems	SDS-PAGE, ESI-MS, MALDI-TOF mass spectrometry; LC–ESI-MS followed by ZIC-HILIC for glycosylated proteins	51,71

ESI-MS, electrospray ionization–mass spectrometry; MALDI-TOF, matrix-assisted laser desorption/ionization–time of flight; DLS, dynamic light scattering; ANSEC, analytical size-exclusion chromatography; ZIC-HILIC, zwitterionic chromatography–hydrophilic interaction liquid chromatography; LC-MS, liquid chromatography–mass spectrometry; SeMet, selenomethionine.

There are two important provisos to the methods and strategies described in this review. First, our experience is dominated by studies with nonmembrane cytosolic and/or fragments of proteins that comprise soluble domains. Second, although the protocols for the 'first attempt' described here have proven to be optimal for the broadest range of proteins, in any individual case, the methods will fail more often than they succeed.

Obtaining the cDNA and creating the expression clone

cDNA. Recently, sequencing efforts and various cDNA consortia have made available large libraries of full-length, sequence-verified cDNAs. Although there are inevitably issues with clone contamination and mix-up, the resources are in general trustworthy. Among the most comprehensive and best annotated is the Mammalian Gene Collection, which maintains a repository of >19,000 human cDNAs, covering ~65% of all annotated genes. For genes or splice variants not easily obtained through more traditional routes, total gene synthesis can be used. Over the past few years, the cost of gene synthesis has dropped almost fivefold, and it will undoubtedly continue to decrease. One advantage of gene synthesis is the ability to change the codon bias of the gene to be more compatible with the recombinant host. However, for *Escherichia coli*, expression strains supplemented with additional tRNAs can often overcome the codon bias of the recombinant gene[17]. For example, in a study of 30 human genes by the Structural Genomics Consortium (SGC), there was no clear advantage in the use of codon-optimized genes compared with the natural sequence expressed in tRNA-supplemented strains (N.A. Burgess-Brown, S. Sharma, F. Sobott, C. Loenarz, U. Oppermann and O. Gileadi; submitted).

Selecting the N and C termini. The objective of recombinant protein expression is usually to produce a sample that supports a certain biochemical or biological activity, such as enzyme catalysis or protein-ligand interactions. Frequently, the desired activity is supported by a discrete domain, and thus it is often not necessary to express the full-length protein to address a particular biological question. In expressing a protein domain, the choice of the N- and C-terminal boundaries represents an important consideration because even small differences can dramatically influence both solubility and expression. For example, Klock and colleagues[18] evaluated a nested set of 2,143 N- and C-terminal truncations from 96 targets and found considerable variation in both solubility and aggregation behavior by altering the protein length by just a few amino acids.

Despite the best efforts, and even for proteins whose domain structure is well-defined, it is not currently possible to predict which specific N- and C-terminal boundaries are most compatible with the expression of a soluble protein. Thus, pragmatism dictates testing many truncated forms of the protein to select one or more for scale-up production. For proteins of known or readily predicted three-dimensional structure, the borders should be engineered to encompass the domain of interest. As an example, ten constructs of the targeted domain might be made at the outset of every project, one corresponding to the full-length protein and nine representing the clones derived from amplifying a combination of three different 5′-end primers and three different 3′-end primers. Gräslund and colleagues have compared the success rate of the nested-primer approach with the predicted success rate if one had chosen only a single 'optimal' construct. In a sample set of 400 human protein domains, the use of multiple constructs increased the probability of generating a soluble protein twofold[19].

To select the sets of PCR primers for proteins with a predictable three-dimensional structure, one should consider prior knowledge of the structure of a related protein, sequence conservation patterns, and predictions of secondary structure or unfolded/disordered regions[20,21]. Widely accepted guidelines are to: (i) remove predicted membrane-spanning regions; (ii) avoid disrupting predicted secondary structural elements; (iii) respect the boundaries of globular domains, if known; and (iv) avoid inclusion of low-complexity regions or hydrophobic residues at the termini[22]. The optimal step size between the nested primers is not yet fully understood; we commonly make constructs to encode proteins that vary in length by 2–10 amino acids at each end[19]. For proteins without a predictable three-dimensional structure, the approximate boundaries of the region of interest might be identified using functional assays and scanning deletion mutagenesis, and then optimal boundaries for expression can be identified using nested sets of PCR primers, as above[23]. Boundaries of structured domains can also be determined experimentally by using limited proteolysis combined with mass spectrometry analysis[24]. Clearly, when using protein fragments, caution should be used in interpreting unexpected biological results.

BOX 1 SUMMARY OF CONSENSUS PROTOCOL

- Obtain the cDNA by amplifying either genomic DNA (prokaryotic genes, or eukaryotic genes with no introns) or full-length, sequence-verified cDNAs (eukaryotes) or by total gene synthesis.
- Use ligation-independent cloning (LIC) to clone the full-length cDNA (or the fragment of interest) into an *E. coli* expression vector.
- Use T7 RNA polymerase–driven expression and an N-terminal oligohistidine tag (include a cleavage site for a sequence-specific protease to enable removal of the tag).
- Express the protein in a derivative of the *E. coli* BL21(DE3) strain, with induction at low temperature (15–25 °C) in rich medium and with good aeration. If expressing proteins from organisms that have codon biases differing from those used by *E. coli*, use a strain supplemented with the appropriate tRNA genes.
- Solubilize and purify the protein in a well-buffered solution containing an ionic strength equivalent to 300–500 mM of a monovalent salt, such as NaCl.
- Use immobilized metal affinity chromatography (IMAC) as the initial purification step.
- If additional purification is required, use size-exclusion chromatography (gel filtration). If necessary, use ion exchange chromatography as a final 'polishing' step.
- The affinity tag may be removed to minimize non-native sequences in the recombinant protein and to achieve further purification. Use a recombinant, hexahistidine-tagged protease and reapply the sample to IMAC column to remove the protease and any cellular proteins that bound to the metal affinity resin.

Cloning

The most common methods now used in our groups to clone target genes into the requisite expression vector rely on homology-based approaches, using either recombination enzymes[25] or ligation-independent cloning (LIC)[26]. Restriction enzyme–based

approaches are used less frequently. A comparison of the methods is shown in **Supplementary Table 1** online.

Recombination-based methods include, for example, the bacteriophage lambda integrase system[27] and the Cre-*lox* recombination system[28]. These methods are rapid, easy and produce few false positives. However, the requirement for special cloning sites imposes constraints: either additional amino acid codons are inserted at either end of the gene, making the PCR primers quite long, or the work-around cloning strategies are more complicated. The unique feature of these methods is the ability to transfer the cloned sequence among a series of compatible vectors that can be used to express the gene in different hosts or with different tags. For bacterial expression, however, the probability of identifying a clone that expresses a soluble protein is increased by making different variants of a single protein in the same *E. coli* host rather than by cloning a single variant into vectors with different tags and expression hosts[19,29].

Ligation-independent cloning, which is used by most of our groups, has the disadvantage compared with recombination-based approaches in that one needs to clone sequences independently into each vector (if this is required). However, the method is inexpensive and simple. One scientist can routinely generate two 96-well plates of distinct clones in a week without the benefit of automation.

Expressing the protein

***E. coli* as the recombinant host for initial studies**. The stably folded, globular domains of prokaryotic and eukaryotic proteins (for example, catalytic domains or protein interaction domains) are a major focus both of the biomedical research community and of our laboratories. These proteins are generally suitable for expression in *E. coli*. Over the years, much effort has been put into optimizing *E. coli* as an expression host for proteins from higher organisms[30]. This strategy has generated a wide arsenal of tools that can be used to increase the yield of soluble protein.

A surprising variety of other classes of proteins, from full-length bacterial and human proteins, to protein complexes, and even some human integral membrane proteins can also be produced in *E. coli*. In terms of full-length proteins, analysis of large-scale protein expression trials shows that up to 50% of proteins from the Eubacteria or Archaea and 10% of proteins from the Eukarya can be expressed in *E. coli* in soluble form[31] (http://targetdb.pdb.org/). Overall, the probability of successfully expressing a soluble protein decreases considerably at molecular weights above ~60 kDa (**Fig. 1**). Proteins that do not express in soluble form may not be modified or folded properly, or may precipitate within *E. coli* through formation of inclusion bodies. Remarkably, expression in a heterologous host does not solely account for the poor success rates; even after extensive screens of expression conditions, 30% of proteins from *E. coli* itself cannot be produced in soluble form when overexpressed in *E. coli*[32].

On the basis of these studies, our view is that the first attempt for the recombinant production of any protein—whatever the source—is to try *E. coli* as the expression host. It is fast and inexpensive to test a wide variety of possible strategies in *E. coli,* and one can complete a fairly comprehensive analysis within a relatively short period of time. Alternative systems should be used only after the *E. coli* system has been reasonably explored. This view balances the fact that there is definitely a lower probability of expressing some classes of proteins in *E. coli* (full-length eukaryotic proteins, integral membrane proteins) compared with other systems (human or insect cells), with the fact that the *E. coli* system is useful in many cases, and also is far more cost-effective and convenient.

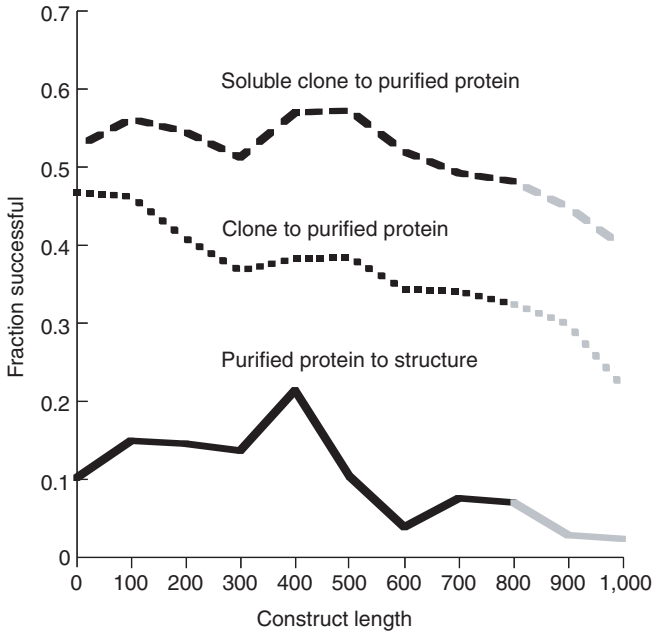

Figure 1 | Solubility as a function of construct length. Fraction of successful purifications and structure determinations as a function of protein length (data from New York Structural GenomiX Research Center). Dotted line, fraction of cloned targets resulting in successful large-scale purifications. Dashed line, fraction of soluble clones (those that express soluble protein at a 1-ml scale) that yield pure protein at large scale. Solid line, fraction of purified proteins resulting in successful crystal structure determinations. There are relatively few targets with lengths greater than 800 amino acids, so these fractions have been extrapolated and are shown in gray.

Strain of *E. coli*. For high-level protein production purposes, BL21(DE3) is an appropriate *E. coli* strain. It has the advantage of being deficient in both *lon* and *ompT* proteases and it is compatible with the T7 *lac*O promoter system[33]. For eukaryotic proteins, it is often important to use BL21(DE3) derivatives carrying additional tRNAs to overcome the effects of codon bias. Historically, ampicillin has been the most commonly used antibiotic-selection marker, but it is being replaced by carbenicillin, which is more stable. Vectors encoding resistance to kanamycin or chloramphenicol are now widely used as well.

Fusion to oligohistidine tags. We suggest that the protein should be produced as a fusion to an affinity tag because tags dramatically aid in protein purification and rarely adversely affect biological or biochemical activity[34]. However, in selecting which tag to use, one is faced with a daunting number of choices. Our groups have explored most of the available options, and we observed that no affinity tag emerged as significantly more efficacious in successfully producing soluble, active recombinant proteins[35]. Despite the lack of a clear winner based on success rate, most of our research groups selected an N-terminal hexahistidine tag that can be removed by a site-specific protease, such as the tobacco etch virus (TEV) protease[36]. However, many other instances can be found in which proteins can be expressed in soluble form only as fusions to other affinity tags[29].

The rationale for the choice of an N-terminal hexahistidine is manifold. First, an N-terminal tag ensures that the bacterial transcription and translation machineries always encounter 5′ and N-terminal sequences that are compatible with robust RNA synthesis and protein expression, respectively. Second, oligohistidine-tagged proteins can be purified using a relatively simple protocol using immobilized metal affinity chromatography (IMAC)[37]. Third, histidine tags rarely affect the characteristics of the protein, which distinguishes it, for example, from glutathione *S*-transferase (GST), which itself is a dimer that then imposes dimerization on the recombinant protein. Fourth, the hexahistidine tag is relatively small and usually does not dramatically alter the solubility properties of the target protein. By contrast, larger tags, such as the maltose-binding protein (MBP), can often increase the apparent solubility of the recombinant moiety, even when the protein is either insoluble by nature, or unstable or unfolded and, therefore, less likely to be active[38–40]. Fifth, for the specific application of protein crystallography, short histidine tags appear to be neutral actors; in most of our projects, we routinely attempt crystallization and NMR structure determination with both cleaved and uncleaved proteins, and their relative representation among the resulting three-dimensional structures is roughly equivalent. A recent PDB-wide survey[41] also indicates that hexahistidine tags do not have a consistent impact on the N-terminal structure of the target protein.

T7 RNA polymerase–based expression vectors. The most commonly used expression systems are based on pET vectors (Merck/EMD; the pET System manual, 2006), which drive expression of a recombinant gene under the control of the T7 RNA polymerase promoter and lac operator[33,42]. The vectors are designed for use in λDE3 lysogen strains of *E. coli*, which harbor a genomic copy of the gene for T7 RNA polymerase under the control of the lac repressor. Under repressive conditions, T7 RNA polymerase is not produced, and transcription of the target gene is negligible. After induction, when the T7 RNA polymerase is produced, most of the cellular protein synthesis machinery will be devoted to producing the target protein. On occasion, low-level expression of T7 polymerase within these strains leads to expression of the recombinant protein and may slow or prevent growth of the transformed bacteria. The expression of such highly toxic proteins can be effected by using T7 lysozyme-expressing strains[42], strains in which the T7 RNA polymerase is under the control of the arabinose promoter[43], by producing the protein in a cell-free system[44] or by driving expression of the recombinant protein directly by the more tightly regulated arabinose promoter system[45].

Expression conditions. Using T7 systems, protein expression can be induced either with the chemical inducer isopropyl-β-D-thiogalactoside (IPTG) or by manipulating the carbon sources during *E. coli* growth (auto-induction; ref. 46 and the pET System manual; Merck/EMD, 2006).

In both cases, the cells can, and should, be grown to high densities (OD_{600} of 4–20) in highly enriched medium[47] in baffled shake flasks[48,49]. Whatever the final cell density, it is advisable to induce the expression of the T7 RNA polymerase at mid-to-late log phase of the growth curve to ensure maximal yield while avoiding the problems associated with cells going into stationary phase (for example, induction of proteases). One feature of the T7 system is that many recombinant proteins often precipitate when expressed at 37 °C, but are soluble when the temperature during induction is 15–25 °C, presumably because slower rates of protein production allow newly transcribed recombinant proteins time to fold properly[50]. Thus, lower temperatures during induction should be used as the default.

Small-scale test expression

Small-scale test expression is widely used as a predictive tool to determine which of the derivative clones actually produces soluble protein and to establish the optimal scale for the large-scale growth. A major concern is that the expression level and solubility of a recombinant protein is influenced by the culture conditions and the degree of aeration, and these parameters do not always scale with culture volume. The results from small and large-scale growth also vary owing to differences in sample preparation and protein purification methods that are used for each scale of growth. Therefore, whereas positive small-scale experiments are often predictive of the results from large-scale growth, there will inevitably be a substantial proportion of false negatives in which an apparently nonexpressed or insoluble protein can be in fact, expressed in soluble form when grown on a larger scale. If the total number of constructs to be tested is small (for example, <20 constructs), it may be wiser to proceed immediately to larger-scale cultures to avoid any potential complications.

For analysis of large numbers of constructs, parallel small-scale protein purification can be performed efficiently in volumes of 1–20 ml, in 96-well format. This scale typically produces 10–200 μg of protein, which is sufficient for many analytical tests. The results can be used to optimize the construct design and experimental conditions before embarking on larger scale purifications[49,51,52].

Protein purification

As a chromatographic procedure, IMAC has the advantages of having strong, specific binding, mild elution conditions and the ability to control selectivity by including low concentrations of imidazole in chromatography buffers. There is a broad array of common resins with slightly different binding capacities and binding strengths, but all tolerate harsh cleaning procedures (TALON Metal Affinity Resins User Manual, Clontech, 2007; the QIAexpressionist, Qiagen, 2003; and HisTrap HP, 1 ml and 5 ml (instructions), Amersham Biosciences, GE Healthcare, 2003). Most purification steps can be integrated by high-performance liquid chromatography; the most commonly used devices are the ÄKTA systems from GE Healthcare.

The final purity of the protein can be optimized by controlling the ratio of recombinant protein to the column size; lower-affinity contaminants can be competed with a relative excess of the histidine-tagged recombinant protein. Accordingly, it is beneficial to determine the amount of the soluble target protein to be loaded on the column, and this can be estimated from small-scale expression trials. As a general rule, to maximize purity, one should load the column with a slight excess over the predicted binding capacity. Although not necessary, it is relatively straightforward to implement these protein purification protocols on automated chromatography systems, which have proven reliable, effective and simple to use.

Preparation of the bacterial lysate. Preparation of the bacterial lysate is a critical step. Optimal conditions maximize cell lysis and the fraction of the recombinant protein that is extracted while minimizing protein oxidation, unwanted proteolysis and sample contamination with genomic DNA. Mechanical lysis by high-pressure homogenization or sonication, or lysis by freeze-thaw procedures

with lysozyme are equivalent in most cases. The lysis buffer should contain a strong buffer (50–100 mM phosphate or HEPES) to overcome the contribution of the bacterial lysate, high ionic strength (equivalent to 300–500 mM NaCl) to enhance protein solubility and stability, protease inhibitors and a reducing agent such as Tris(2-carboxyethyl) phosphine hydrochloride (TCEP) to prevent oxidation of the protein. Loading large amounts of bacterial lysate (>1 l culture volume) on small (<1 ml) affinity columns may require prior removal of any particulate or viscous material. This can be accomplished by using enzymes that degrade nucleic acid and cell-wall material, such as DNase or Benzonase (Merck/EMD) and lysozyme, respectively. Some of the enzymes used in lysis are less active in the presence of reducing agents or high salt concentration; optimal lysis may require sequential addition of the components. Clarified lysates can also be filtered before loading on the affinity columns.

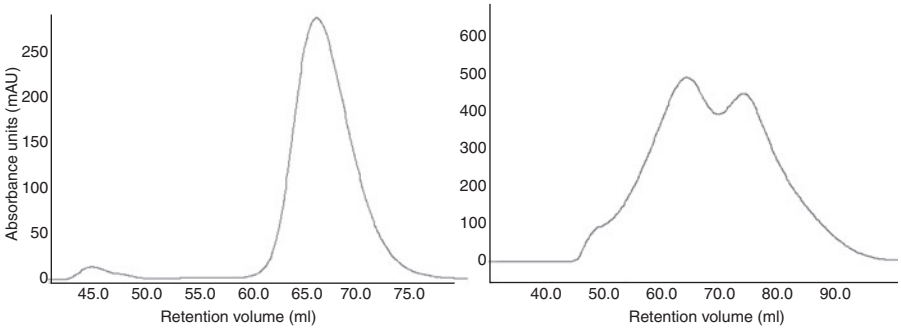

Figure 2 | Gel filtration profiles. Representative good (left) and bad (right) gel-filtration profiles of two different proteins purified on an ÄKTAxpress system using a HiLoad Superdex 200 column (GE Healthcare).

IMAC purification is performed in phosphate buffer, pH 8.0 and an ionic strength equivalent to 300–500 mM NaCl. HEPES buffer (and, to a lesser extent, Tris buffer) at pH 7.5–8.0 can also be used. It has been consistently observed that conditions of high ionic strength (for example, 500 mM NaCl) maintain solubility and stability of the widest variety of proteins. Indeed, a substantial fraction of proteins precipitate if the salt concentration is reduced to physiological levels, particularly as the protein becomes more pure and concentrated. The choice of NaCl as the salt is mainly historical and, although not systematically explored, there is no reason to believe that sodium and chloride are optimal. Indeed, sodium and chloride levels in the cell are very low and are probably never the physiologically relevant counter-ions for intracellular proteins. A modest amount of imidazole (see resin manufacturer's recommendations) should be included in the cell extraction buffer to reduce binding of less histidine-rich proteins to the IMAC column. For intracellular proteins, care should be taken to maintain a reducing environment. TCEP, unlike dithiothreitol (DDT), is compatible with all known IMAC matrices. Finally, inclusion of glycerol (10%) during protein purification enhances the solubility and stability of many proteins.

Chromatography. After the lysate is loaded on the IMAC column, it should be washed with buffer including an intermediate concentration of imidazole (see manufacturer's instructions), which will elute weakly bound contaminants without sacrificing large amounts of the recombinant protein. It is sometimes necessary to optimize the wash step with respect to the concentration of imidazole as well as the volume of the wash. Finally, the recombinant protein should be with a step gradient (for example, 300 mM imidazole). If EDTA and DTT are added after IMAC; add the EDTA first to sequester any nickel that has leached off and that could react with the DTT.

The choice of gel filtration as the next step may be surprising, considering its lower resolving power compared with ion exchange or other adsorption chromatography methods, but this step is often sufficient after IMAC if the protein was abundant in the lysate. Moreover, gel filtration is more generic, can be performed in any buffer condition, and can be used to resolve the oligomerization state of the target protein. In some cases, if the protein is judged insufficiently pure for the intended purpose, one can remove the tag with a histidine-tagged TEV protease and perform IMAC again as an additional 'generic' purification step, collecting the recombinant protein in the flowthrough. This step very efficiently removes histidine-rich proteins derived from the expression host, which may have copurified in the primary IMAC procedure, as well as the cleaved tag and the histidine-tagged protease.

Protein characterization

Characterizing the purified protein in some detail reduces the risk of wasting resources on protein material of inadequate quality. It also provides a means to ensure that different batches of the same protein have similar properties. Below, we outline a simple, generic protein characterization protocol that allows the experimentalist to judge whether the correct protein has been purified, whether additional molecular species are present and to estimate the approximate protein concentration. Other characterization methods that are very informative but not as widely applied, such as mass spectrometry, static or dynamic light scattering, and measuring protein thermal stability, are described in **Supplementary Methods**.

Inspection of gel filtration chromatogram. If size exclusion chromatography was used as the last purification step, a close look at the chromatogram is essential. Symmetric elution profiles are characteristic of homogeneous proteins, whereas asymmetric profiles reflect inhomogeneous, or partially aggregated, samples (**Fig. 2**), or whether the column itself is in poor condition. The elution profiles will also reveal the primary oligomerization state. The presence of additional oligomerization states may be of biological significance, or may be a sign of nonspecific aggregation. If the protein elutes in the void volume of the chromatogram, the protein is most likely forming large, nonspecific aggregates, which may be an indication of improper folding and compromised activity. It is also of value to analyze individual peaks by SDS-PAGE or mass spectrometry to analyze the protein in each peak.

SDS-PAGE analysis. After protein purification, samples should be resolved by denaturing SDS-PAGE. If stained with a dye such as Coomassie brilliant blue, the intensity of the bands will usually be proportional to the amount of protein[53]. This allows the purity of the sample to be estimated and whether the purified protein is of the expected size.

Table 3 | Common *E. coli* proteins that copurify by IMAC

Protein	Accession number	Monomer mass (kDa)
GroES	NP_418566	10.39
Fur	NP_415209	16.79
SlyD	NP_755987	20.85
CA	NP_414668	25.10
RplB	P60422	29.86
DnaJ	NP_414556	41.10
GroEL	AAS75782	57.35
DnaK	NP_414555	69.11

Based on information in references 54 and 55, and on unpublished observations.

UV absorption spectroscopy. To quantify the amount and concentration of purified protein, the simplest and most common method is the Bradford assay[53], which measures the binding of Coomassie brilliant blue to the protein. As some proteins bind the dye anomalously, it is also useful to measure the UV absorption at A_{280} and calculate the concentration of the protein by using the predicted molar extinction coefficient at A_{280} (http://www.expasy.org/tools/protparam.html). By taking a UV absorption spectrum, it is also possible to uncover contamination with DNA or RNA, or reveal common copurifying cofactors (for example, NAD, FAD, heme).

Storing purified protein. Aliquots of the protein to be stored should be placed in thin-walled PCR plastic tubes, frozen in liquid nitrogen and stored at –80 °C. Small aliquots should be frozen to avoid damaging freeze-thaw cycles, and aliquots should be thawed on ice. Concentrated proteins (for example, >1 mg/ml) tend to be more stable to freeze-thaw cycles. Proteins are usually concentrated using centrifuge-driven filter devices with adequate molecular weight size cutoffs. Care should be taken during centrifugation to avoid local over-concentration and irreversible precipitation or aggregation of the protein on the filtration membrane.

It is advisable to explore the stability of the protein to concentration and freeze-thaw cycles before processing the entire batch. The frozen and thawed sample should be compared with protein that was not frozen for biochemical activity, visible precipitation, changes in physical properties (for example, dynamic light scattering or gel filtration profile) or crystallization characteristics. In our collective experience, relatively few proteins are irreversibly inactivated by one freeze-thaw cycle. In those rare instances, the protein can be stored at 4 °C for short periods of time, at –20 °C in high concentrations of glycerol, or as an ammonium sulfate suspension.

Common 'traps' and 'pitfalls'

Poor lysis. In small-scale test expression and solubility trials designed to assess the extent to which a protein partitions to the soluble or insoluble fractions, it is important to ensure that the cells are lysed and fractionated properly. Although this is not technically challenging, we have found that it is very common to fail to achieve complete bacterial lysis, which leads to an underestimation of the proportion of recombinant protein in the soluble fraction. Care should also be taken when removing the soluble fraction after centrifugation; it is relatively easy to contaminate the soluble fraction with insoluble material, which can lead to an overestimate of the amount of recombinant protein in the soluble fraction. As a quality control, it is advisable to inspect the protein profiles of the fractions using SDS gel electrophoresis. Some cellular proteins characteristically resolve into the soluble and insoluble fractions and these serve as excellent internal controls (**Supplementary Fig. 1** and **2** online).

The recombinant protein fails to bind the IMAC column. The pH of the lysate should be 7.5–8.0 for efficient binding, and the buffer should not contain chelators (EDTA or citrate), high imidazole concentrations (for example, >30 mM for Ni-NTA resins) or DTT. In some instances, it is necessary to reduce the amount of imidazole in the loading buffer to <5 mM. The column must be properly charged with metal ions and, when charging columns, make sure the concentrated $NiSO_4$ solution is buffered and set to pH 7.5. It is also important to remember that imidazole is a base; the final solutions must be adjusted to the correct pH. In some cases the target protein may bind weakly to the IMAC column, so the concentration of imidazole in the wash step should be reduced (for example, 20 mM).

The wrong or a mutant protein was expressed or purified. An incorrect protein may occasionally be expressed and purified, which most commonly results from a simple clone mix-up. In that instance the problem will be detected either by gel electrophoresis or mass spectrometry of the purified protein.

If the recombinant protein is expressed at low levels, it is also relatively common to purify an endogenous *E. coli* protein that binds to, and elutes from, the IMAC column and that also adventitiously migrates with the predicted mobility of the target protein[54]. In some cases, this *E. coli* protein may even appear to be induced after the expression of T7 RNA polymerase. Determining whether you have purified your recombinant protein or an endogenous bacterial protein can readily be accomplished with mass spectrometry, but is more difficult by denaturing gel electrophoresis. A western blot to the affinity tag can sometimes be useful to track the recombinant protein.

If the expression construct is sequenced before the experiment, errors introduced in primer synthesis or PCR will be detected. In practice, PCR-generated sequence errors are so rare that it is often more practical to do the expression trials first, and to sequence the successful expression constructs later. Of course, if none of the constructs express a protein, it is essential to sequence the expression clones and, ultimately, to sequence the clones selected for scale-up and purification.

Bacterial proteins copurify with the recombinant protein. Copurification of *E. coli* proteins with the histidine-tagged recombinant protein is very common, especially when the expression level of the recombinant protein is low. Contaminants include proteins that contain multiple histidine residues (for example, SlyD; **Table 3**), and molecular chaperones that may bind to the resin directly or to the recombinant protein[54,55]. The affinity resin has limited capacity, so loading near-saturating amounts of the recombinant protein on a column improves purity. Tag cleavage followed by affinity purification is also effective in removing contaminants, as these proteins are unaffected by the protease and bind to the column after reapplication of the cleavage reaction. Samples copurifying with chaperones should be regarded with suspicion because this indicates that the protein may have some unfolded character. In cases where the target protein cannot be separated from the chaperones by additional chromatography, use an alternative expression system, process a different construct of the protein or try working with a closely related ortholog.

Samples contain additional proteins or multiple protein species or states. If the protein target is contaminated with other proteins, one can perform additional purification steps such as ion-exchange chromatography. Purifying samples contaminated with different post-translationally modified species or proteolytic fragments of the same protein is more challenging, but not necessarily intractable. For example, different phosphorylated states of a protein can sometimes be resolved using ion-exchange chromatography[56].

'Pure' samples precipitate or fail to concentrate. Pure proteins often precipitate out of solution, even at relatively low (<1 mg/ml) concentrations. This behavior is sometimes coupled with sample inhomogeneity, either in the form of contaminating protein or alternate folded states. Precipitation can also occur by aggregation owing to the presence of hydrophobic or hydrophilic patches on the surface of the target protein. In either case, the problem worsens as the protein concentration increases. There are no generic solutions but some potential solutions, which must be explored for each protein, are to: find a more stabilizing buffer through screening using analytical gel filtration or thermal denaturation (see **Supplementary Methods**), maintain the protein at lower concentration (<0.1–0.5 mg/ml), maintain an adequate reduced state to prevent protein oxidation (>5 mM DTT, refreshed as required), maintain the salt concentration at high levels (ionic strength >500 mM of a monovalent salt), add glycerol to 10%, add arginine in the range of 50–500 mM, add a mild nondenaturing detergent (0.1% β-octylglucoside) or keep the protein at its optimal temperature (determined empirically).

Rescue strategies

In even the best of circumstances, it is unusual to generate a soluble version of any given protein on the first attempt. As such, it is important to have a series of alternative approaches. Here we provide various suggestions in the order in which we would usually apply them.

Changing expression conditions. Adjustment of the expression conditions seldom results in radical changes but, as some optimization can be done quite easily, it is worth the effort. The first step is to lower the temperature to slow down protein production. Different types of media can also be tested; rich media, such as Terrific Broth, 2×YT or ZYP5052 (auto-induction), often support good expression. Changing the *E. coli* strain can also improve expression of a soluble protein[51].

Expression of more variants of the protein sequence. As described above, it is important to test the expression of a range of constructs to identify those that express a soluble derivative. We suggest expressing as many as 10 constructs in the initial attempts. If this proves unsuccessful, then it may be advisable to explore additional constructs, particularly if one has knowledge that a structurally related protein can be expressed in soluble form.

Alternate tags. Our consensus strategy is to append an N-terminal histidine tag to each construct. If the histidine-tagged recombinant protein does not express or is insoluble, then the probability that it will be expressed in an active form with another N-terminal fusion partner is reduced considerably. Our advice, therefore, is not to iteratively append different N-terminal fusions but to first explore a C-terminal fusion to the histidine tag instead. Some proteins that are completely insoluble with an N-terminal histidine tag can be expressed in soluble form with a C-terminal histidine tag[57].

Although we do not advise extensive sampling of other N-terminal fusions, this strategy can sometimes lead to production of soluble, stable fusion protein. If the aim is to study the function of the target protein, and the fusion protein is an acceptable reagent, then it may be an appropriate strategy. However, this approach has its caveats. In the absence of a robust and quantitative functional assay, one reasonably uses solubility as a proxy for function. However, proteins that are soluble only with a larger tag can be 'dragged' into solution by the tag, and revert to an insoluble form if the fusion partner is removed[38–40].This indicates that the integrity of the recombinant protein as a fusion protein may be suspect. For example, wild-type GFP is mostly insoluble when expressed in *E. coli* at 37 °C but is largely expressed in the soluble fraction as an MBP fusion[58]. Nonetheless, bacterial colonies expressing the MBP-GFP fusions display only weak fluorescence, suggesting that the GFP is non-functional (G.S. Waldo; unpublished data). Accordingly, before any functional studies, considerable attention should be paid to whether a target protein appears to be soluble only because it is a passenger on a larger tag.

Coexpression of interacting proteins. Many proteins are obligate components of multiprotein assemblies and these often require an interacting protein for correct folding and stability[21,59,60]. Such proteins, and those with unstructured polypeptide chain segments, often cannot be expressed in *E. coli* in soluble form, but it has proven possible to improve the properties of these proteins by coexpressing the cognate interacting protein[61–63]. This strategy is only starting to be used in the large-scale projects, in those cases when entire families of interacting proteins are being studied.

Ligand supplementation. Many proteins can be stabilized by the binding of a small molecule—a principle that has found widespread application in generic screening for protein ligands[64,65]. This property can be exploited to increase the proportion of recombinant protein expressed in soluble form or to stabilize a protein during purification. If a sufficiently soluble, cell-permeable and avid ligand is available, one can use it to stabilize newly synthesized proteins and promote solubility[66,67]. This concept has also not yet been explored sufficiently in a systematic way.

Other expression hosts. If bacterial expression is unsuccessful to this point, other hosts should be considered. Common eukaryotic alternatives are the baculovirus expression system in insect cells[68], the yeasts *Pichia pastoris*[69] and *Saccharomyces cerevisiae*[70], human cells[71], or cell-free systems using prokaryotic or eukaryotic extracts[72–76]. These cell-free systems, which have been used extensively to generate thousands of purified proteins for structural studies[77–79], can be used to produce proteins that are toxic to *E. coli*[79] and can use PCR-amplified linear DNA fragments, without cloning into a vector, for screening and optimization.

All these other expression systems are reasonably simple to use, but they are somewhat more time-consuming to work with than are bacteria and require equipment less commonly found in a typical laboratory.

Coexpression of chaperones. Proper *in vivo* folding of a recombinant protein can be promoted by coexpression of molecular chaperones, which are typically produced from cotransformed plasmids carrying several chaperones with synergistic effects, such

as the pG-Tf2 vector[80]—a combination of GroEL-GroES[81] and trigger factor[82]. In our hands, chaperones have been used successfully only in isolated cases, and we know of no study of considerable size that has demonstrated broad efficacy.

Refolding. A commonly tried but only episodically successful protocol to rescue insoluble protein is to denature the protein and try to refold it *in vitro*. The method can be successful[83,84], particularly for extracellular proteins. However, even the most robust protocols only refold a small fraction of the input protein, and it is difficult to purify the refolded fraction. The best procedures use an activity assay to monitor refolding, and affinity reagents that select any refolded, active protein. We would advise using refolding as a last resort for intracellular proteins.

Summary

The methods and strategies for protein expression and purification have been reviewed for the expert many times in excellent, comprehensive ways. Here we attempted to provide a resource for those entering the field, reflecting the experiences of our groups in the application of the various methods to large numbers of proteins. We understand there are many possible routes to obtain high-quality protein and acknowledge that the methods described above should be considered as a starting point that can be embellished once sufficient expertise has been obtained. Detailed protocols for the methods described in this review can be found in the **Supplementary Methods**.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

The Structural Genomics Consortium is a registered charity (number 1097737) that receives funds from the Canadian Institutes for Health Research, the Canadian Foundation for Innovation, Genome Canada through the Ontario Genomics Institute, GlaxoSmithKline, Karolinska Institutet, the Knut and Alice Wallenberg Foundation, the Ontario Innovation Trust, the Ontario Ministry for Research and Innovation, Merck & Co., Inc., the Novartis Research Foundation, the Swedish Agency for Innovation Systems, the Swedish Foundation for Strategic Research and the Wellcome Trust. The New York Structural GenomiX Research Center for Structural Genomics is supported by the US National Institute of General Medical Sciences (U54 GM074945). Work at the MDC was supported by the German Federal Ministry for Education and Research (BMBF) through the Leitprojektverbund Proteinstrukturfabrik and the German National Genome Network (NGFN; FKZ 01GR0471, 01GR0472), and by the Fonds der Chemischen Industrie. The Protein Sample Production Facility is funded by the Helmholtz Association of German Research Centres. The China Structural Genomics Consortium is supported by the National 863 Hi-Tech Research and Development Program of China. The Israel Structural Proteomics Center is supported by The Israel Ministry of Science, Culture and Sport, the Divadol Foundation, the Neuman Foundation, the European Commission Sixth Framework Research and Technological Development Programme 'SPINE2-Complexes' Project under contract 031220. The RIKEN Structural Genomics/Proteomics Initiative was supported by the National Project on Protein Structural and Functional Analyses, Ministry of Education, Culture, Sports, Science and Technology of Japan. The Joint Center for Structural Genomics is supported by the US National Institutes of Health (NIH) Protein Structure Initiative grant U54 GM074898 from the NIGMS. The Northeast Structural Genomics Consortium is supported by the NIH NIGMS (U54-GM074958). The Midwest Center for Structural Genomics is supported by the NIH (GM074942) and by the US Department of Energy, Office of Biological and Environmental Research (DE-AC02-06CH11357). The Oxford Protein Production Facility is funded by the UK Medical Research Council and Biotechnology and Biological Sciences Research Council. SPINE2-Complexes is funded by the European Commission (contract 031220) under the Framework 6 RTD Programme and is coordinated from the Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Oxford, UK . The Berkeley Structural Genomics Center is supported by the NIH (GM62412). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIGMS or the NIH.

1. Sauder, J.M. *et al.* High throughput protein production and crystallization at NYSGXRC. in *Structural Proteomics: High-Throughput Methods* Vol. 426 (eds. B. Kobe, M.Guss & H. Thomas) 561–575 (Humana Press, Totowa, New Jersey, USA, 2008).
2. Büssow, K. *et al.* Structural genomics of human proteins-target selection and generation of a public catalogue of expression clones. *Microb. Cell Fact.* **4**, 21 (2005).
3. Heinemann, U., Büssow, K., Mueller, U. & Umbach, P. Facilities and methods for the high-throughput crystal structural analysis of human proteins. *Accounts Chem. Res.* **36**, 157–163 (2003).
4. Banci, L. *et al.* First steps towards effective methods in exploiting high-throughput technologies for the determination of human protein structures of high biomedical value. *Acta Crystallogr.* **D62**, 1208–1217 (2006).
5. Aricescu, A.R. *et al.* Eukaryotic expression: developments for structural proteomics. *Acta Crystallogr.* **D62**, 1114–1124 (2006).
6. Heinemann, U. Establishing a structural genomics platform: The Berlin-based Protein Structure Factory. *Gene Funct. Dis.* **3**, 25–32 (2002).
7. Scheich, C., Kummel, D., Soumailakakis, D., Heinemann, U. & Büssow, K. Vectors for co-expression of an unrestricted number of proteins. *Nucleic Acids Res.* **35**, e43 (2007).
8. Bartlam, M., Xu, Y. & Rao, Z. Structural proteomics of the SARS coronavirus: a model response to emerging infectious diseases. *J. Struct. Funct. Genomics* **8**, 85–97 (2007).
9. Gong, W.M. *et al.* Structural genomics efforts at the Chinese Academy of Sciences and Peking University. *J. Struct. Funct. Genomics* **4**, 137–139 (2003).
10. Albeck, S. *et al.* Three-dimensional structure determination of proteins related to human health in their functional context at The Israel Structural Proteomics Center (ISPC). *Acta Crystallogr.* **D61**, 1364–1372 (2005).
11. Lesley, S.A. *et al.* Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. *Proc. Natl. Acad. Sci. USA* **99**, 11664–11669 (2002).
12. Montelione, G.T., Zheng, D., Huang, Y.J., Gunsalus, K.C. & Szyperski, T. Protein NMR spectroscopy in structural genomics. *Nat. Struct. Biol.* **7** (Suppl.), 982–985 (2000).
13. Acton, T.B. *et al.* Robotic cloning and Protein Production Platform of the Northeast Structural Genomics Consortium. *Methods Enzymol.* **394**, 210–243 (2005).
14. Alzari, P.M. *et al.* Implementation of semi-automated cloning and prokaryotic expression screening: the impact of SPINE. *Acta Crystallogr.* **D62**, 1103–1113 (2006).
15. Gileadi, O. The scientific impact of the Structural Genomics Consortium: a protein family and ligand-centered approach to medically-relevant human proteins. *J. Struct. Funct. Genomics* **8**, 107–119 (2007).
16. Gileadi, O. *et al.* Methods in Molecular Biology. in *Structural Proteomics: High-Throughput Methods.* Vol. 426 (eds., B. Kobe, M. Guss & T. Huber) 222–246 (Humana Press, Totowa, New Jersey, USA, 2008).
17. You, J., Cohen, R.E. & Pickart, C.M. Construct for high-level expression and low misincorporation of lysine for arginine during expression of pET-encoded eukaryotic proteins in *Escherichia coli*. *Biotechniques* **27**, 950–954 (1999).
18. Klock, H.E., Koesema, E.J., Knuth, M.W. & Lesley, S.A. Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins* published online, doi: 10.1002/prot.21786 (14 November 2007).
19. Gräslund, S. *et al.* The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. *Protein Expr. Purif.* (in the press).
20. Ginalski, K., Elofsson, A., Fischer, D. & Rychlewski, L. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015–1018 (2003).
21. Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F. & Jones, D.T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–2139 (2004).
22. Yang, Z.R., Thomson, R., McNeil, P. & Esnouf, R.M. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **21**, 3369–3376 (2005).
23. Cornvik, T. *et al.* An efficient and generic strategy for producing soluble human proteins and domains in *E. coli* by screening construct libraries. *Proteins* **65**, 266–273 (2006).
24. Gao, X. *et al.* High-throughput limited proteolysis/mass spectrometry for protein domain elucidation. *J. Struct. Funct. Genomics* **6**, 129–134 (2005).
25. Hartley, J.L., Temple, G.F. & Brasch, M.A. DNA cloning using *in vitro* site-

specific recombination. *Genome Res.* **10**, 1788–1795 (2000).
26. Aslanidis, C. & de Jong, P.J. Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res.* **18**, 6069–6074 (1990).
27. Landy, A. Dynamic, structural, and regulatory aspects of lambda site-specific recombination. *Annu. Rev. Biochem.* **58**, 913–949 (1989).
28. Guo, F., Gopaul, D.N. & Van Duyne, G.D. Asymmetric DNA bending in the *Cre-lox*P site-specific recombination synapse. *Proc. Natl. Acad. Sci. USA* **96**, 7143–7148 (1999).
29. Hammarström, M., Hellgren, N., van Den Berg, S., Berglund, H. & Härd, T. Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Sci.* **11**, 313–321 (2002).
30. Peti, W. & Page, R. Strategies to maximize heterologous protein expression in *Escherichia coli* with minimal cost. *Protein Expr. Purif.* **51**, 1–10 (2007).
31. Braun, P. & LaBaer, J. High throughput protein production for functional proteomics. *Trends Biotechnol.* **21**, 383–388 (2003).
32. Vincentelli, R. *et al.* Medium-scale structural genomics: strategies for protein expression and crystallization. *Accounts Chem. Res.* **36**, 165–172 (2003).
33. Studier, F.W., Rosenberg, A.H., Dunn, J.J. & Dubendorff, J.W. Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol.* **185**, 60–89 (1990).
34. Uhlén, M., Forsberg, G., Moks, T., Hartmanis, M. & Nilsson, B. Fusion proteins in biotechnology. *Curr. Opin. Biotechnol.* **3**, 363–369 (1992).
35. Arnau, J., Lauritzen, C., Petersen, G.E. & Pedersen, J. Current strategies for the use of affinity tags and tag removal for the purification of recombinant proteins. *Protein Expr. Purif.* **48**, 1–13 (2006).
36. Carrington, J.C. & Dougherty, W.G. A viral cleavage site cassette: identification of amino acid sequences required for tobacco etch virus polyprotein processing. *Proc. Natl. Acad. Sci. USA* **85**, 3391–3395 (1988).
37. Porath, J. Immobilized metal ion affinity chromatography. *Protein Expr. Purif.* **3**, 263–281 (1992).
38. Nallamsetty, S. & Waugh, D. Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners. *Protein Expr. Purif.* **45**, 175–182 (2006).
39. Nallamsetty, S. & Waugh, D.S. A generic protocol for the expression and purification of recombinant proteins in *Escherichia coli* using a combinatorial His_6-maltose binding protein fusion tag. *Nat. Protoc.* **2**, 383–391 (2007).
40. Waugh, D.S. Making the most of affinity tags. *Trends Biotechnol.* **23**, 316–320 (2005).
41. Carson, M., Johnson, D.H., McDonald, H., Brouillette, C. & Delucas, L.J. His-tag impact on structure. *Acta Crystallogr.* **63**, 295–301 (2007).
42. Dubendorff, J.W. & Studier, F.W. Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with *lac* repressor. *J. Mol. Biol.* **219**, 45–59 (1991).
43. Wycuff, D.R. & Matthews, K.S. Generation of an AraC-araBAD promoter-regulated T7 expression system. *Anal. Biochem.* **277**, 67–73 (2000).
44. Shimizu, Y. *et al.* Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19**, 751–755 (2001).
45. Guzman, L.M., Belin, D., Carson, M.J. & Beckwith, J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose pBAD promoter. *J. Bacteriol.* **177**, 4121–4130 (1995).
46. Studier, F.W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).
47. Lesley, S.A. High-throughput proteomics: protein expression and purification in the postgenomic world. *Protein Expr. Purif.* **22**, 159–164 (2001).
48. Tunac, J. A new high-aeration capacity shake-flask system. *J. Ferm. Bioeng.* **68**, 15–159 (1989).
49. Brodsky, O. & Cronin, C.N. Economical parallel protein expression screening and scale-up in *Escherichia coli*. *J. Struct. Funct. Genomics* **7**, 101–108 (2006).
50. Vera, A., Gonzalez-Montalban, N., Aris, A. & Villaverde, A. The conformational quality of insoluble recombinant proteins is enhanced at low growth temperatures. *Biotechnol. Bioeng.* **96**, 1101–1106 (2007).
51. Berrow, N.S. *et al.* Recombinant protein expression and solubility screening in *Escherichia coli*: a comparative study. *Acta Crystallogr.* **D62**, 1218–1226 (2006).
52. Page, R. *et al.* Scalable high-throughput micro-expression device for recombinant proteins. *Biotechniques* **37**, 364–370 (2004).
53. Bradford, M.M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–254 (1976).
54. Bolanos-Garcia, V.M. & Davies, O.R. Structural analysis and classification of native proteins from *E. coli* commonly co-purified by immobilised metal affinity chromatography. *Biochim. Biophys. Acta* **1760**, 1304–1313 (2006).
55. Howell, J.M., Winstone, T.L., Coorssen, J.R. & Turner, R.J. An evaluation of in vitro protein-protein interaction techniques: assessing contaminating background proteins. *Proteomics* **6**, 2050–2069 (2006).
56. Bullock, A.N., Debreczeni, J., Amos, A.L., Knapp, S. & Turk, B.E. Structure and substrate specificity of the Pim-1 kinase. *J. Biol. Chem.* **280**, 41675–41682 (2005).
57. Alam, M., Ho, S., Vance, D.E. & Lehner, R. Heterologous expression, purification, and characterization of human triacylglycerol hydrolase. *Protein Expr. Purif.* **24**, 33–42 (2002).
58. Kapust, R.B. & Waugh, D.S. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* **8**, 1668–1674 (1999).
59. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z. & Dunker, A.K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584 (2002).
60. Frenkiel-Krispin, D. *et al.* Plant transformation by *Agrobacterium tumefaciens*: modulation of single-stranded DNA-VirE2 complex assembly by VirE1. *J. Biol. Chem.* **282**, 3458–3464 (2007).
61. Tolia, N.H. & Joshua-Tor, L. Strategies for protein coexpression in *Escherichia coli*. *Nat. Methods* **3**, 55–64 (2006).
62. Romier, C. *et al.* Co-expression of protein complexes in prokaryotic and eukaryotic hosts: experimental procedures, database tracking and case studies. *Acta Crystallogr.* **62**, 1232–1242 (2006).
63. Bullock, A.N., Debreczeni, J.E., Edwards, A.M., Sundström, M. & Knapp, S. Crystal structure of the SOCS2-elongin C-elongin B complex defines a prototypical SOCS box ubiquitin ligase. *Proc. Natl. Acad. Sci. USA* **103**, 7637–7642 (2006).
64. Vedadi, M. *et al.* Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proc. Natl. Acad. Sci. USA* **103**, 15835–15840 (2006).
65. Niesen, F.H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat. Protoc.* **2**, 2212–2221 (2007).
66. Elleby, B. *et al.* High-level production and optimization of monodispersity of 11beta-hydroxysteroid dehydrogenase type 1. *Biochim. Biophys. Acta* **1700**, 199–207 (2004).
67. Strauss, A. *et al.* Improved expression of kinases in Baculovirus-infected insect cells upon addition of specific kinase inhibitors to the culture helpful for structural studies. *Protein Expr. Purif.* **56**, 167–176 (2007).
68. Smith, G.E., Summers, M.D. & Fraser, M.J. Production of human beta interferon in insect cells infected with a baculovirus expression vector. *Mol. Cell. Biol.* **3**, 2156–2165 (1983).
69. Boettner, M., Prinz, B., Holz, C., Stahl, U. & Lang, C. High-throughput screening for expression of heterologous proteins in the yeast *Pichia pastoris*. *J. Biotechnol.* **99**, 51–62 (2002).
70. Holz, C., Hesse, O., Bolotina, N., Stahl, U. & Lang, C. A micro-scale process for high-throughput expression of cDNAs in the yeast *Saccharomyces cerevisiae*. *Protein Expr. Purif.* **25**, 372–378 (2002).
71. Aricescu, A.R., Lu, W. & Jones, E.Y. A time- and cost-efficient system for high-level protein production in mammalian cells. *Acta Crystallogr.* **D62**, 1243–1250 (2006).
72. Yokoyama, S. Protein expression systems for structural genomics and proteomics. *Curr. Opin. Chem. Biol.* **7**, 39–43 (2003).
73. Kigawa, T. *et al.* Preparation of *Escherichia coli* cell extract for highly productive cell-free protein expression. *J. Struct. Funct. Genomics* **5**, 63–68 (2004).
74. Matsuda, T. *et al.* Cell-free synthesis of zinc-binding proteins. *J. Struct. Funct. Genomics* **7**, 93–100 (2006).
75. Endo, Y. & Sawasaki, T. Cell-free expression systems for eukaryotic protein production. *Curr. Opin. Biotechnol.* **17**, 373–380 (2006).
76. Mikami, S., Masutani, M., Sonenberg, N., Yokoyama, S. & Imataka, H. An efficient mammalian cell-free translation system supplemented with translation factors. *Protein Expr. Purif.* **46**, 348–357 (2006).
77. Yokoyama, S., Terwilliger, T.C., Kuramitsu, S., Moras, D. & Sussman, J.L. RIKEN aids international structural genomics efforts. *Nature* **445**, 21 (2007).
78. Murayama, K. *et al.* Crystal structure of the rac activator, Asef, reveals its autoinhibitory mechanism. *J. Biol. Chem.* **282**, 4238–4242 (2007).
79. Miyazono, K. *et al.* Novel protein fold discovered in the PabI family of restriction enzymes. *Nucleic Acids Res.* **35**, 1908–1918 (2007).
80. Nishihara, K., Kanemori, M., Yanagi, H. & Yura, T. Overexpression of trigger factor prevents aggregation of recombinant proteins in *Escherichia coli*. *Appl. Environ. Microbiol.* **66**, 884–889 (2000).
81. Wynn, R.M., Song, J.L. & Chuang, D.T. GroEL/GroES promote dissociation/reassociation cycles of a heterodimeric intermediate during alpha(2)beta(2) protein assembly. Iterative annealing at the quaternary structure level. *J. Biol. Chem.* **275**, 2786–2794 (2000).
82. Kaiser, C.M. *et al.* Real-time observation of trigger factor function on translating ribosomes. *Nature* **444**, 455–460 (2006).

83. Willis, M.S. *et al.* Investigation of protein refolding using a fractional factorial screen: a study of reagent effects and interactions. *Protein Sci.* **14**, 1818–1826 (2005).
84. Vincentelli, R. *et al.* High-throughput automated refolding screening of inclusion bodies. *Protein Sci.* **13**, 2782–2792 (2004).
85. Kim, S.H. *et al.* Structural genomics of minimal organisms and protein fold space. *J. Struct. Funct. Genomics* **6**, 63–70 (2005).
86. Lesley, S.A. & Wilson, I.A. Protein production and crystallization at the joint center for structural genomics. *J. Struct. Funct. Genomics* **6**, 71–79 (2005).
87. Kreusch, A. & Lesley, S.A. High throughput cloning, expression and purification technologies. in *Genomics, Proteomics, and Vaccines* (ed., G. Grandi) 171–184 (Wiley Press, Chichester, UK, 2004).
88. McMullan, D. *et al.* High-throughput protein production for X-ray crystallography and use of size exclusion chromatography to validate or refute computational biological unit predictions. *J. Struct. Funct. Genomics* **6**, 135–141 (2005).
89. Stols, L., Millard, C.S., Dementieva, I. & Donnelly, M.I. Production of selenomethionine-labeled proteins in two-liter plastic bottles for structure determination. *J. Struct. Funct. Genomics* **5**, 95–102 (2004).
90. Geerlof, A. *et al.* The impact of protein characterization in structural proteomics. *Acta Crystallogr.* **D62**, 1125–1136 (2006).

The authors are:

Susanne Gräslund[1], Pär Nordlund[1], Johan Weigelt[1], James Bray[2], Opher Gileadi[2], Stefan Knapp[2], Udo Oppermann[2], Cheryl Arrowsmith[3], Raymond Hui[3], Jinrong Ming[3], Sirano dhe-Paganon[3], Hee-won Park[3], Alexei Savchenko[3], Adelinda Yee[3], Aled Edwards[3], Renaud Vincentelli[4], Christian Cambillau[4], Rosalind Kim[5], Sung-Hou Kim[5], Zihe Rao[6], Yunyu Shi[7], Thomas C Terwilliger[8], Chang-Yub Kim[8], Li-Wei Hung[8], Geoffrey S Waldo[8], Yoav Peleg[9], Shira Albeck[9], Tamar Unger[9], Orly Dym[9], Jaime Prilusky[9], Joel L Sussman[9], Ray C Stevens[10], Scott A Lesley[10,11], Ian A Wilson[10,11], Andrzej Joachimiak[12], Frank Collart[12], Irina Dementieva[12], Mark I Donnelly[12], William H Eschenfeldt[12], Youngchang Kim[12], Lucy Stols[12], Ruying Wu[12], Min Zhou[12], Stephen K Burley[13], J Spencer Emtage[13], J Michael Sauder[13], Devon Thompson[13], Kevin Bain[13], John Luz[13], Tarun Gheyi[13], Fred Zhang[13], Shane Atwell[13], Steven C Almo[14], Jeffrey B Bonanno[14], Andras Fiser[14], Sivasubramanian Swaminathan[15], F William Studier[15], Mark R Chance[16], Andrej Sali[17], Thomas B Acton[18], Rong Xiao[18], Li Zhao[18], Li Chung Ma[18], John F Hunt[19], Liang Tong[19], Kellie Cunningham[18], Masayori Inouye[18], Stephen Anderson[18], Heleema Janjua[18], Ritu Shastry[18], Chi Kent Ho[18], Dongyan Wang[18], Huang Wang[18], Mei Jiang[18], Gaetano T Montelione[18], David I Stuart[20,23], Raymond J Owens[20,23], Susan Daenke[20,23], Anja Schütz[21], Udo Heinemann[21], Shigeyuki Yokoyama[22], Konrad Büssow[21,24], Kristin C Gunsalus[18,24]