Design of high-throughput methods of protein production for structural biology

Raymond C Stevens

Address: Departments of Molecular Biology and Chemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, SR101, La Jolla, CA 92037, USA.
E-mail: stevens@scripps.edu

Structure 2000, **8**:R177–R185

0969-2126/00/$ – see front matter

Introduction

High-throughput protein expression and purification has a central, pivotal role in structural genomics. In fact, crystallographic-quality protein production on the scale required to generate tens to hundreds of different proteins per day will probably be the greatest obstacle for the conversion of protein structure determination to a high-throughput format. High-throughput efforts in structural biology place unique restrictions on protein expression and purification. First, the majority of gene constructs must be expressed in a synchronous fashion. Second, the purification protocol applied to the majority of expressed proteins must be as similar as possible and produce the very high-quality material that is needed for structural studies. The first requirement can be addressed by employing either large or small N-terminal expression tags, and the second hurdle can be circumvented with the use of affinity purification tags. However, the drawback in incorporating affinity tags in crystallization studies is that in many cases these tags introduce flexible portions to the protein of interest that are not conducive to crystallization or lead to various forms of microheterogeneity. Protease cleavage sites allow removal of these flexible tag regions, but conditions often have to be optimized for each reaction, requiring fine-tuned processing to be incorporated in a high-throughput environment. Given the above caveats on current technology, however, affinity-tag systems are still the most useful to date given the restrictions placed on high-throughput methods.

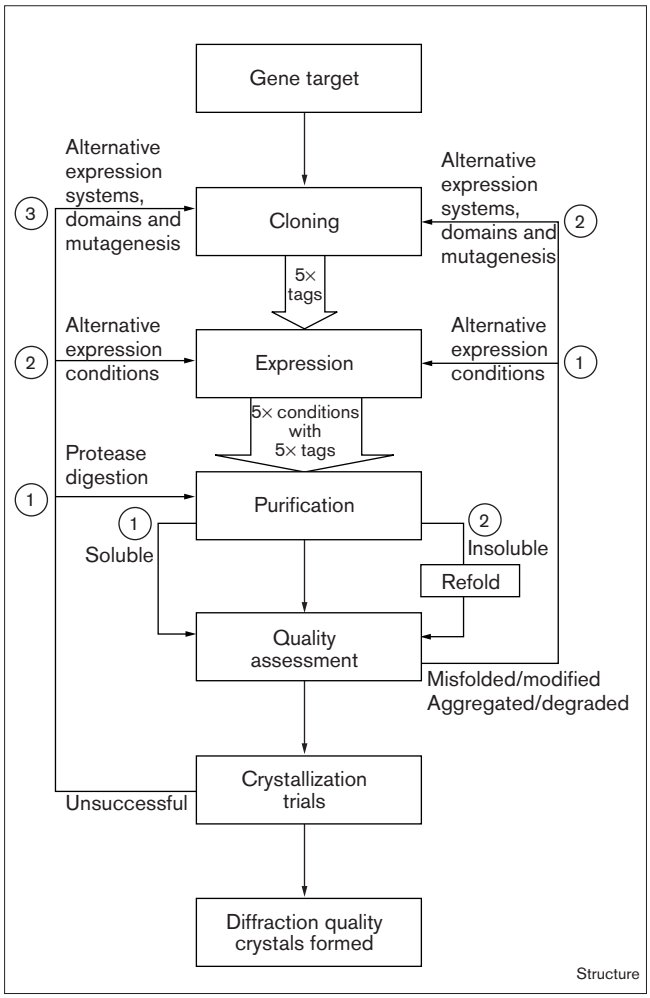
Several groups and commercial companies are currently designing high-throughput protein production, in particular, systems based on *Escherichia coli* are described in this review. In addition, other systems are being investigated, including *in vitro* expression systems [1,2], methods that employ baculovirus infection of insect cells with a peptide tag for antipeptide monoclonal antibody purification [3], and yeast-based expression using a single tag [4] or dual tags [5]. A distinct advantage of *in vitro* expression systems is the facilitation of selenomethionine (SeMet) incorporation or ^{15}N labeling. Scale-up has been a problem with this technology, however, and is currently being addressed [1,2]. Because all of these systems are still in the early stages of development, data collection in a systematic and comprehensive manner will be necessary to arrive at the most efficient and cost-effective solution for high-throughput protein expression and purification.

The parallel production of protein targets, expressed under a range of conditions with a variety of affinity tags, will result in a number of successful conditions for protein production, along with many unsuccessful conditions. By analyzing which conditions produce viable samples for a given type of protein and which conditions do not work, a large knowledge base will be produced. This knowledge base will provide the foundation for a statistically relevant method of predicting effective expression conditions for protein samples with similar physical and biochemical properties. As this knowledge base grows in volume, the predictive power should improve, resulting in a larger number of successes and an exponential growth of the knowledge base itself. Through this iterative process, trends in protein production will become apparent, and the speed and efficiency of high-throughput protein production will be greatly enhanced. Each aspect from cloning to protein quality control is discussed here with an emphasis towards high-throughput methods and collection of success/failure data in a systematic manner.

Cloning

Recombinant protein yield and solubility are highly dependent on the specific protein sequence, as well as on the vector, host cell, and culture conditions used. For optimal efficiency, various combinations should be simultaneously screened to determine the conditions that yield the 'best' protein, for example, choosing five different expression clones for each protein of interest (see Figure 1). Cloning using restriction enzymes typically cannot be used for high-throughput approaches, owing to the complication of selecting compatible and appropriate restriction enzymes for each cloning procedure. Additionally, multiple steps of experimental refinement and treatment must be performed using the restriction enzyme(s) of interest. High-throughput cloning therefore requires procedures based on the polymerase chain reaction (PCR). A first step involves the design of specific gene PCR amplification primers, followed by screening of potential PCR-amplified clones for proper insert orientation. Sequence analysis of positive clones must be performed, to confirm that a proper reading frame has been obtained and that no PCR-introduced errors are present. This step is followed by a final archiving of selected plasmid DNA samples. Automated systems are available for colony

Figure 1



Overview of the steps involved in high-throughput protein production for structural studies. With the correct positive and negative controls, one should be able to pass through such a flowchart only once in order to successfully obtain protein samples suitable for structural analysis. The numbers indicate preferences in the flowchart diagram.

picking, gridding, and microarraying (e.g. the 'Q Pix' system sold by Genetix [http://www.genetix.co.uk] and the Gene Suite™ available from GeneMachines [http://www.genemachines.com]) and also for sequencing (e.g. fluorescence-based systems available from Perkin Elmer Biosystems, Amersham Pharmacia Biotech, and Visible Genetics).

To generate expression vector clones, cloning systems such as the Invitrogen Echo™ system [6], the Gibco/Life Technologies Gateway™ system, or the Novagen pTriEx-1 cloning system [7] may be advantageous. These generic processes streamline the expression cloning process by alleviating costly, time-consuming recloning steps and avoiding the use of restriction enzymes in the cloning and subcloning process. For example, the Invitrogen system inserts the PCR-amplified gene fragment into a TOPO™ 'donor' vector using a topoisomerase-I-adapted plasmid. CRE recombinase is then used in a second step to introduce the correct cloned sequence into a loxP-adapted recipient vector (*E. coli*, insect, yeast, and mammalian choices are all available) for subsequent protein expression studies.

Expression

High-throughput approaches will rely upon both prokaryotic and eukaryotic hosts. Bacterial expression systems are advantageous for a variety of reasons, most notably that protein overexpression is usually obtained without any post-translational modification heterogeneity. In addition, *E. coli* protein expression is cheaper and faster than eukaryotic systems. In comparison, more expensive and slower eukaryotic systems will be necessary for the expression of some subsets of proteins, especially those that require post-translational modifications for proper folding and activity. Currently, many companies have available a variety of vector/host expression systems (see the Supplementary material section). Numerous heterologous gene expression systems are available (see the Supplementary material section), and additional variants can be constructed using combinations of strong promoters and tight regulators (as listed in Table 1) along with the proper transcription initiation and translation signals. These systems produce recombinant gene products in an efficient and regulatable manner [8,9]. Cost-effective *E. coli* expression, using T7 RNA polymerase plus the T7 promoter [10,11] with induction for high-yield recombinant protein overexpression, appears to be appropriate for 'first pass' efforts.

As previously mentioned, there is an important advantage in using tags in high-throughput protein expression and purification efforts, so that all proteins will have a generic 'handle'. Several N-terminal expression tags are available (see Table 2), ranging from large tags (e.g. *E. coli* thioredoxin [12], *Schistosoma japonicum* glutathione-*S*-transferase [GST] [13] and *E. coli* maltose-binding protein [MBP] [14]) down to fairly small tags (e.g. S-tag [15], His-tag [16] and T7-tag) [17–23]. Although expression data and correlation to crystallization have not been conducted in a thorough manner, it is generally believed that the expression fusion tags need to be removed, particularly for large fusion partners. Table 3 summarizes the possible choices for incorporating tag-cleavage and self-cleavage into expression constructs. Unfortunately, the endopeptidases suffer from many limitations, including the presence of peptide secondary cleavage site activity (leading to proteolytically damaged products), incomplete sample cleavage (leading to product heterogeneity which hampers crystallization), and inhibition of cleavage by properly folded proteins (requiring partial denaturation for successful fusion-tail cleavage) [24]. The viral proteases to date have proven to be the most selective and useful for structural biology studies. There is the additional possibility of utilizing

Table 1

Control elements used for *E. coli* recombinant protein expression.

Promoter	Regulation	Inducer	Level of expression	Comment
T7 bacteriophage	*lacI*^{q}	IPTG	Very high	Utilizes T7 RNA polymerase. High-level inducible overexpression commonly obtained. T7lac system for tight control of induction needed for more toxic clones. Expensive induction.
trc (hybrid) *E. coli*	*lacI, lacI*^{q}	IPTG	Moderately high	Lower level expression versus T7 systems, but high-level, regulated expression still possible. Expensive induction.
p_L (λ)	λcIts857	Temperature shift to 42°C	Moderately high	Temperature-sensitive host required. Less likelihood of 'leaky' uninduced expression [80].
araBAD	*araC*	L-Arabinose	Variable, from high to low level	Can fine-tune expression levels in a dose-dependent manner (tight regulation possible). Inexpensive inducer.

expression constructs that incorporate different combinations of tags, for multiple affinity purification procedures that allow for increased selectivity of purification [8,17,18,21,25–27].

Optimizing expression levels

High-throughput expression requires the parallel induction of all clones in one expression run under the same conditions, which requires the presence of N-terminal expression tags to standardize baseline recombinant protein expression levels. In order to enable reliable prediction of expression constructs that generate soluble or insoluble expressed proteins, empirical trials need to be performed, altering expression conditions (e.g. the temperature or inducer concentration used for a run) and observing the solubilities and stabilities of the recombinant proteins, that are obtained [28]. As outlined in Figure 1, a probable combination of five different conditions should be probed. Prescreening using SDS–PAGE, in combination with Western blot analysis, is advantageous to analyze expression constructs on a small-scale and to determine the levels of proteins produced. This screening also provides information on the degradation or aggregation of proteins from potential expression clones.

Host strain genotype is important for obtaining optimal expression levels, and many specialized *E. coli* strains have been developed. For example, the BL21 *lon* and *ompT* protease-deficient strain improves the likelihood of isolating intact full-length recombinant proteins [10]. Another important variable is the induction level used [29]. Although high-level expression can usually be obtained, optimizing the yield of properly folded protein might require a reduction in the induction level. Finally, media formulations must be considered. In some instances, optimal expression might require supplementary salts (e.g. Zn^{2+}, Cu^{2+}, Ca^{2+}, Fe^{2+}, Fe^{3+}) and potential cofactors or prosthetic groups (e.g. heme, FAD, FMN, tetrahydrobiopterin) to be included [30]; for example, for proteins with catalytically or structurally important metal centers expression trials should contain metal ions in the culture media [31,32]. Studies should determine optimal media formulations for maximal recombinant protein yield; for example, with rich media formulations more active proteins are sometimes produced [33]. Current media conditions are old and outdated, and were developed for different applications. There is a need to develop new media conditions optimized for high-expression level protein production under high-density growth conditions.

Purification

Recombinant protein purification is facilitated by the use of high-yield expression systems so that the desired protein is produced in an enriched form. Purification is further simplified by the presence of affinity purification tags. For high-throughput processing, initial efforts have focused on soluble and insoluble purification strategies using the cost-effective His-tag [34]. The incorporation of a His-tag allows for a generic single-step purification using nickel-nitrilotriacetate or other immobilized metal-affinity chromatography resins. In practice, a second purification step is almost always required, using either ion-exchange or size-exclusion chromatography. Equipment originally designed for high-throughput DNA manipulations, such as 96-well robotics manifolds, are available for parallel protein sample purification from Qiagen (http://www.qiagen.com).

A final factor to consider, once purified proteins are obtained, is that protein samples for structural studies must be of fairly high concentration, although the range varies from 1 mg/ml to greater than 60 mg/ml (average values are 10–15 mg/ml). Because all proteins vary in their tendency to aggregate and precipitate from solution, it is impossible to define one specific concentration or buffer condition for all protein targets. However, the rules for

Table 2

Fusion tags used for recombinant protein expression and purification.

Tag	Size	Fusion tag location	Tag type	Comments
His-tag	6, 8 or 10 aa	N, C, internal	Purification	Most common purification tag used for immobilized metal-affinity chromatography (IMAC) one-step purification [81]. Purification possible even under denaturing conditions [82]. Tag possibly influences crystallization.
T7-tag	11 or 16 aa	N, internal	Purification, enhanced expression	Monoclonal antibody-based purification (denaturing low pH elution needed). Leaves unnatural N-terminal amino acids on the recombinant protein. Possibly enhanced expression levels as the T7-tag is derived from the T7 gene 10, which is the naturally most abundant phage T7 gene product.
S-tag	15 aa	N, C, internal	Purification and detection	S-protein (104 aa, ribonuclease A minus S-tag peptide sequence) modified resin affinity purification. RNAse S assay possible for quantitative assay of expression levels.
FLAG™ peptide (DYKDDDDK)	8 aa	N, C	Purification	Ca^{2+}-dependent monoclonal antibody purification with EDTA elution. Tag cleavable with enterokinase [83].
Thioredoxin	109 aa (11.7 kDa)	N, C	Purification and enhanced expression	Affinity purification with phenylarsine oxide-modified (ThioBond) resin.
His-patch thioredoxin	109 aa (11.7 kDa)	N, C	Purification and enhanced expression	Use of His-patch modified thioredoxin for IMAC purification [84].
lacZ (β-galactosidase)	116 kDa	N, C	Purification	Purification using *p*-amino-phenyl-β-D-thiogalactoside-modified sepharose. Classical tag used for protecting peptides from proteolytic degradation. Fusion proteins with this tag have a high tendency to be insoluble. Active enzyme is a tetramer.
Chloramphenicol acetyltransferase	24 kDa	N	Secretion, purification and detection	Chloramphenicol–sepharose purification. Enzymatic assay possible for quantitation.
trpE	27 kDa	N	Purification	Often form insoluble precipitates. Hydrophobic interaction chromatography purification.
Avidin/streptavidin/ *Strep*-tag			Purification and secretion	Biotin affinity purification and streptavidin affinity purification (*Strep*-tag) [85].
T7gene10	260 aa	N	Purification and enhanced expression	Produces insoluble fusion protein (potential enhanced expression for toxic clones).
Staphylococcal protein A	14 kDa (or 31 kDa)	N	Purification and secretion	IgG antibody affinity purification possible (denaturing low pH elution needed). Fusion protein secretion due to protein A signal sequence [86].
Streptococcal protein G	28 kDa	N, C	Purification and secretion	Albumin affinity purification, low pH elution needed. Fusion protein secretion due to protein G signal sequence.
Glutathione-*S*-transferase (GST)	26 kDa	N	Purification	Glutathione affinity or GST antibody purification. Enzymatic activity assay possible for quantitative analysis. Fusion proteins form dimers.
Dihydrofolate reductase (DHFR)	25 kDa	N	Purification	Methotrexate-linked agarose used for purification.
Cellulose-binding domains (CBP)	156 aa/ 114 aa/ 107 aa/	N N C	Purification and secretion	Cellulose-based resins used for affinity purification with water elution [87,88]. Different constructs available for cytoplasmic or periplasmic expression. Fusion proteins susceptible to proteolysis between the fusion partners [89].
Maltose-binding protein (MBP)	40 kDa	N, C	Purification and secretion	Amylose affinity purification with maltose elution.
Galactose-binding protein			Purification	Galactose-sepharose purification.
Calmodulin-binding protein (CBP)	4 kDa	N, C	Purification and detection	Calmodulin/Ca^{2+} affinity purification with EDTA elution. Can potentially assay expression levels with ^{32}P-cAMP kinase.

Table 2 continued

Tag	Size	Fusion tag location	Tag type	Comments
Hemagglutinin influenza virus (HAI)			Purification	
Green fluorescent protein	220 aa	N, C	Detection	Used as reporter gene fusion for detection purposes [90]. Used at one time for possible refolding tag.
HSV-tag	11 aa	C	Purification	Monoclonal antibody-based purification (denaturing low pH elution needed).
B-tag (VP7 protein region of bluetongue virus)			Purification	Anti-BTag antibody purification.
Polyarginine	5–15 aa	C	Purification	S-sepharose (cationic resin) purification. Fusion proteins potentially insoluble.
Polycysteine	4 aa	N	Purification	Thiopropyl-sepharose purification.
Polyphenylalanine	11 aa	N	Purification	Phenyl-superose (hydrophobic interaction chromatography) purification.
$(Ala\text{-}Trp\text{-}Trp\text{-}Pro)_n$			Purification	
Polyaspartic acid	5–16 aa	C	Purification	Anionic resin purification.
KSI	125 aa	N	Enhanced expression	High-level inclusion body production.
c-myc			Purification	Anti-myc antibody purification.
OmpT/OmpA /PelB /DsbA/DsbC	22 aa/21 aa /20 aa /208 aa/236 aa	N	Secretion	Periplasmic leader sequences for potential protein export and folding [91], as well as potential disulfide bond formation and isomerization.
Chitin-binding domain		N, C	Expression	Used in the Impact™ system, with intein-based expression constructs.
NusA	495 aa	N	Possible enhanced solubility	Potentially improve solubility for proteins that are overexpressed.
Ubiquitin	76 aa	N	Possible enhanced solubility	Ubiquitin fusions observed to increase *E. coli* expressed recombinant protein solubility.
lac repressor			Purification	*lac* operator affinity purification.
T4 gp55				
Growth hormone, N terminus				

approximating the required concentration are relatively simple. First, the protein concentration should be in a range such that crystallization factorial conditions will allow the protein to reach the solution solubility limit. Second, there must be sufficient protein present in an aliquoted droplet to generate a crystal large enough for X-ray diffraction analysis. For example, a 40 nl droplet can maximally produce a single protein crystal 50 μm in size, assuming a 50% solvent content in the crystal and a starting protein concentration of 10 mg/ml. Hence, a protein-concentrating step must normally be used in protein purifications for crystallization studies. Ion-exchange chromatography has proven useful as a second 'polishing' purification step, as well as directly yielding concentrated solutions upon column elution.

The use of inclusion bodies and refolding screens

Many heterologous proteins that are overexpressed lead to the formation of insoluble protein aggregates known as inclusion bodies [23,30]. Current estimates are that 15–20% of human gene constructs expressed in *E. coli* are soluble, 20–40% form as inclusion bodies, and the remainder do not express significantly or are degraded. Inclusion-body formation can be minimized, or potentially avoided, by adopting a combinatorial approach towards the expression of soluble recombinant protein [10,35,36]. Several modifications in expression conditions should be incorporated into *in vivo* refolding screens to attempt to minimize unfavorable solution behavior: screening for the best host/expression construct combinations; induction at lower temperatures [37,38]; lowering the concentration of inducing agent; altering the media composition (e.g. adding sucrose or polyols [39]); co-expressing chaperones and other *in vivo* folding enhancers [40–45]; expressing the protein as a fusion with a 'solubilizing partner' such as GST, MBP, thioredoxin [46] or NusA [47,48]. In addition, periplasmic secretion can influence recombinant protein solubility [49–52], making leader sequence tags a consideration for high-throughput

Table 3

Cleavage sites used in recombinant protein expression and purification.

Excision site (↓)	Cleavage enzyme/self-cleavage	Comments
Asp-Asp-Asp-Asp-Lys↓	Enterokinase	The site will not cleave if followed by a proline residue. Secondary cleavage sites at other basic residues, depending on conformation of protein substrate. Active from pH 4.5 to 9.5 and between 4°C and 45°C [24].
Ile-Glu/Asp-Gly-Arg↓	Factor Xa protease	Will not cleave if followed by proline and arginine. Secondary cleavage sites following Gly-Arg sequences.
Leu-Val-Pro-Arg↓Gly–Ser	Thrombin	Secondary cleavage sites. Biotinylated form available for removal with immobilized streptavidin.
Glu-Asn-Leu-Tyr-Phe-Gln↓Gly	TEV protease	Seven-residue recognition site, making it a highly site-specific protease. Active over a wide range of temperatures. Protease available as a His-tag fusion protein, allowing for protease removal after recombinant protein cleavage.
Leu-Glu-Val-Leu-Phe-Gln↓Gly-Pro	PreScission™ protease	Genetically engineered form of human rhinovirus 3C protease with a GST fusion tag, allowing for facile cleavage and purification of GST-tagged proteins along with protease removal after recombinant protein cleavage. Enables low-temperature cleavage of fusion proteins containing the eight-residue recognition sequence.
Specific intein-encoded sequences	Intein 1 and intein 2	Uses self-cleavable affinity tags. Even after cleavage unnatural termini are present on the protein of interest.
Signal sequences	Signal peptidases	Cleavage of leader sequences concomitant with protein export from the cytoplasm.

recombinant protein expression trials. Finally, an initial report on developing a mutation screen for soluble protein using green fluorescent protein (GFP) chimeric fusions has been proposed [53].

Inclusion-body-derived proteins can be advantageous for high-throughput methods. They are produced in high yields, and are generally protected from proteolytic degradation. Properly folded proteins can be produced from these intractable inclusion bodies using a variety of solubilization and refolding schemes [43,54–62]. The use of inclusion-body routes is especially attractive for the production of proteins that are toxic to host cells. Unfortunately, high-throughput refolding screens must still be developed, especially as different proteins have highly variable refolding efficiencies [20]. In addition, partial success at *in vitro* refolding has been obtained with hydrophobic interaction chromatography [23,63], but this method has only been marginally tested. Recently, a factorial refolding screen was reported by Gouaux and coworkers [64], but their current system requires an activity assay to assess the success of the refolding procedure. Their system can easily be modified to monitor the success of refolding using fluorescence or circular dichroism spectroscopy [43]. One attractive format for querying protein solubility behavior uses a microdrop screening method to optimize solvent conditions for nuclear magnetic resonance (NMR) spectroscopy of proteins [65]; this methodology might be used to develop high-throughput refolding screens. Unfortunately, current refolding screens are of the hit-and-miss variety [24]. Current methods do not provide rational refolding pathways, allowing for the formation of complex mixtures when attempts are made to renature completely denatured polypeptide chains into properly folded proteins. During generic refolding processes, severe heterogeneity problems can arise owing to the formation of multiple refolded species and aggregates. Hence, efforts aimed at recombinant protein refolding should probe as many factors as possible to maximize the probability of obtaining properly folded native conformation protein.

Quality assessment

Protein crystal growth requires stringent protein purity. Samples must be pure in terms of lacking contaminants, but must also be relatively monodisperse [66], homogeneous, and 'conformationally pure', lacking denatured species and other structural microheterogeneities that adversely affect crystal growth [67–69]. Microheterogeneities, such as variations in primary, secondary, tertiary, and quarternary structure, as well as the presence of various aggregation states, also affect purification and crystallization properties. Every protein is unique in terms of physical properties such as hydrophobic content, surface charge, and solubility behavior with the addition of precipitants, making the propensity of each protein to crystallize very protein-specific. Some proteins can crystallize with impurities or microheterogeneities present, whereas others cannot. It is generally accepted, however, that for ease in crystallization one should have a protein that is as pure and homogeneous as possible.

Prior to the performance of crystallization trials, the purity and homogeneity of protein samples must be confirmed. Automated high-throughput SDS–PAGE and matrix-assisted laser desorption ionization (MALDI) mass spectrometry [70,71] can be used for purity assessment, with dynamic light scattering measurements [66,72–74] used to verify sample dispersity and the degree of aggregation. Sampling miniaturization issues are already being addressed, for example, high-throughput MALDI mass spectroscopy [75] and automated 96-well format bioanalysis procedures [76,77] are being developed in recent proteomics efforts. The ability to screen and analyze results for numerous crystallization trials will allow the correlation of the above measured properties as a function of the processing variables, allowing for improvements in future protein production efforts.

Conclusions

A tremendous number of anecdotal stories exist regarding successes and failures of protein expression, purification, and crystallization of macromolecules. For example, the influence of affinity tags on recombinant protein structure and stability, and the use of additives in media growth have been much debated. This type of information in the Protein Data Bank is not complete. Most expression studies are not performed in a systematic fashion, and failure data is rarely reported. These deficiencies require that further cataloguing and characterization of the influence of different process variables be included in high-throughput purification procedures.

Numerous attractive possibilities currently exist for high-throughput protein production efforts. Current attempts focus on systematic methods for manipulating and processing expressed clones in parallel. It is naive, however, to believe that one simple, global solution will be possible. It will be important to approach the development of high-throughput protein production efforts from a 'learning system' viewpoint. Systematic studies will lay the groundwork for the assembly of comprehensive information databases that can then be used to refine the procedures necessary for efficient genome-scale protein expression and purification efforts [78]. Experience gained will guide later efforts and will assist in solving problems identified in the first rounds of protein production. Finally the problems encountered in the production of more difficult protein targets, such as integral and membrane-associated proteins, still need to be addressed [79].

Supplementary material

Supplementary material including the URLs for several companies that produce high-throughput expression systems is available at http://current-biology.com/supmat/supmatin.htm.

Acknowledgements

I greatly appreciate interacting with Peter Schultz, Scott Lesley and Marc Nasoff at the Genomics Institute for the Novartis Research Foundation and Mark Knuth and Ron Swanson at Syrrx, Inc., where much of this work is currently being pursued. I also thank Jim Graziano, Chris Lee, Mike Hanson and Marianne Patch for helpful suggestions on this manuscript.

References

1. Kigawa, T., Yabuki, T. & Yokoyama, S. (1999). Large-scale protein preparation using the cell-free synthesis. *Tanpakushitsu Kakusan Koso* **44**, 598-605.
2. Kigawa, T., *et al.*, & Yokoyama, S. (1999). Cell-free production and stable-isotope labeling of milligram quantities of proteins. *FEBS Lett.* **442**, 15-19.
3. Albala, J.S. & Humphery-Smith, I. (1999). Array-based proteomics: high-throughput expression and purification of IMAGE consortium cDNA clones. *Curr. Opin. Mol. Ther.* **1**, 680-684.
4. Schuster, M., *et al.*, & Werner, G. (2000). Protein expression strategies for identification of novel target proteins. *J. Biomol. Screening* **5**, 89-97.
5. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. & Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030-1032.
6. Liu, Q., Li, M.Z., Leibham, D., Cortez, D. & Elledge, S.J. (1998). The univector plasmid–fusion system, a method for rapid construction of recombinant DNA without restriction enzymes. *Curr. Biol.* **8**, 1300-1309.
7. Novy, R., Yaeger, K., Monsma, S. & Scott, M. (1999). pTriEx-1 multisystem vector for protein expression in *E. coli*, mammalian, and insect cells. *inNOVAtions* **10**, 1-5.
8. Makrides, S.C. (1996). Strategies for achieving high-level expression of genes in *Escherichia coli. Microbiol. Rev.* **60**, 512-538.
9. Fernandez, J.M. & Hoeffler, J.P. (1999). *Gene Expression Systems: Using Nature for the Art of Expression.* Academic Press, San Diego, CA.
10. Studier, F.W. & Moffatt, B.A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189**, 113-130.
11. Studier, F.W., Rosenberg, A.H., Dunn, J.J. & Dubendorff, J.W. (1990). Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol.* **185**, 60-89.
12. LaVallie, E.R., DiBlasio, E.A., Kovacic, S., Grant, K.L., Schendel, P.F. & McCoy, J.M. (1993). A thioredoxin gene fusion expression system that circumvents inclusion body formation in the *E. coli* cytoplasm. *BioTechnology* **11**, 187-193.
13. Smith, D.B. & Johnson, K.S. (1988). Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione-*S*-transferase. *Gene* **67**, 31-40.
14. di Guan, C., Li, P., Riggs, P.D. & Inouye, H. (1988). Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein. *Gene* **67**, 21-30.
15. Kim, J.-S. & Raines, R.T. (1993). Ribonuclease S-peptide as a carrier in fusion proteins. *Protein Sci.* **2**, 348-356.
16. Hochuli, E., Bannwarth, W., Dobeli, H., Gentz, R. & Stuber, D. (1988). Genetic approach to facilitate purification of recombinant proteins with a novel metal chelate adsorbent. *BioTechnology* **6**, 1321-1325.
17. Uhlen, M. & Moks, T. (1990). Gene fusions for purposes of expression: an introduction. *Methods Enzymol.* **185**, 129-143.
18. Ford, C.F., Suominen, I. & Glatz, C.E. (1991). Fusion tails for the recovery and purification of recombinant proteins. *Protein Express. Purif.* **2**, 95-107.
19. Nilsson, B., Forsberg, G., Moks, T., Hartmanis, M. & Uhlen, M. (1992). Fusion proteins in biotechnology and structural biology. *Curr. Opin. Struct. Biol.* **2**, 569-575.
20. LaVallie, E.R. & McCoy, J.M. (1995). Gene fusion expression systems in *Escherichia coli. Curr. Opin. Biotechnol.* **6**, 501-506.
21. Nilsson, J., Stahl, S., Lundeberg, J., Uhlen, M. & Nygren, P.-A. (1997). Affinity fusion strategies for detection, purification, and immobilization of recombinant proteins. *Protein Express. Purif.* **11**, 1-16.
22. Hannig, G. & Makrides, S.C. (1998). Strategies for optimizing heterologous protein expression in *Escherichia coli. Trends Biotechnol.* **16**, 54-60.
23. Berne, P.F., Doublie, S. & Carter, C.W., Jr. (1999). Molecular biology for structural biology. In *Crystallization of Nucleic Acids and Proteins. A Practical Approach*. (Ducruix, A. & Giege, R., eds), pp. 45-73, Oxford University Press, Oxford.
24. LaVallie, E.R., McCoy, J.M., Smith, D.B. & Riggs, P. (1994). Enzymatic and chemical cleavage of fusion proteins. In *Current Protocols in Molecular Biology*. pp. 16.4.5-16.4.17, John Wiley and Sons, Inc, New York, NY.
25. Kim, J.-S. & Raines, R.T. (1994). Peptide tags for a dual affinity fusion system. *Anal. Biochem.* **219**, 165-166.

26. Müller, K.M., Arndt, K.M., Bauer, K. & Plückthun, A. (1998). Tandem immobilized metal-ion affinity chromatography/immunoaffinity purification of His-tagged proteins – evaluation of two anti-His-tag monoclonal antibodies. *Anal. Biochem.* **259**, 54-61.
27. Cocca, B.A., Seal, S.N. & Radic, M.Z. (1999). Tandem affinity tags for the purification of bivalent anti-DNA single-chain Fv expressed in *Escherichia coli*. *Protein Express. Purif.* **17**, 290-298.
28. Riggs, P., LaVallie, R. & McCoy, J.M. (1994). Introduction to expression by fusion protein vectors. In *Current Protocols in Molecular Biology*. pp. 16.4.1-16.4-4, John Wiley and Sons, Inc, New York, NY.
29. Meerman, H.J. & Georgiou, G. (1994). Construction and characterization of a set of *E. coli* strains deficient in all known loci affecting the proteolytic stability of secreted recombinant proteins. *BioTechnology* **12**, 1107-1110.
30. Mitraki, A. & King, J. (1989). Protein folding intermediates and inclusion body formation. *BioTechnol.* **7**, 690-697.
31. Goodwill, K., Sabatier, C. & Stevens, R.C. (1998). The crystal structure of tyrosine hydroxylase with bound co-factor analog and iron at 2.3 Å resolution: the self-hydroxylation of Phe300 and the pterin binding site. *Biochemistry* **37**, 13437-13445.
32. Hanson, M.A. & Stevens, R.C. (2000). Cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 Å resolution. *Nat. Struct. Biol.* **7**, 687-692.
33. Moore, J.T., Arvinder, U., Maley, F. & Maley, G.F. (1993). Overcoming inclusion body formation in a high level expression system. *Protein Express. Purif.* **4**, 160-163.
34. Arnold, F.H. (1991). Metal-affinity separations: a new dimension in protein processing. *BioTechnology* **9**, 151-156.
35. Wall, J.G. & Pluckthun, A. (1995). Effects of overexpressing folding modulators on the *in vivo* folding of heterologous proteins in *Escherichia coli*. *Curr. Opin. Biotechnol.* **6**, 507-516.
36. Georgiou, G. & Valax, P. (1996). Expression of correctly folded proteins in *Escherichia coli*. *Curr. Opin. Biotechnol.* **7**, 190-197.
37. Burton, N., Cavallini, B., Kanno, M., Moncollin, V. & Egly, J.-M. (1991). Expression in *Escherichia coli*: purification and properties of the yeast general transcription factor TFIID. *Protein Express. Purif.* **2**, 432-441.
38. Schein, C.H. (1989). Production of soluble recombinant proteins in bacteria. *BioTechnol.* **7**, 1141-1147.
39. Blackwell, J.R. & Horgan, R. (1991). A novel strategy for production of a highly expressed recombinant protein in an active form. *FEBS Lett.* **295**, 10-12.
40. Goloubinoff, P., Gatenby, A.A. & Lorimer, G.H. (1989). GroE heat-shock proteins promote assembly of foreign prokaryotic ribulose bisphosphate carboxylase oligomers in *Escherichia coli*. *Nature* **337**, 44-47.
41. Wynn, R.M., Davie, J.R., Cox, R.P. & Chuang, D.T. (1992). Chaperonins GroEL and GroES promote assembly of heterotetramers ($\alpha_2\beta_2$) of mammalian mitochondrial branched-chain α-keto acid decarboxylase in *Escherichia coli*. *J. Biol. Chem.* **267**, 12400-12403.
42. Lee, S.C. & Olins, P.O. (1992). Effect of overproduction of heat shock chaperones GroESL and DnaK on human procollagenase production in *Escherichia coli*. *J. Biol. Chem.* **267**, 2849-2852.
43. Rudolph, R. & Lilie, H. (1996). *In vitro* folding of inclusion body proteins. *FASEB J.* **10**, 49-56.
44. Thomas, J.G. & Baneyx, F. (1996). Protein misfolding and inclusion body formation in recombinant *Escherichia coli* cells overexpressing heat-shock proteins. *J. Biol. Chem.* **271**, 11141-11147.
45. Hayhurst, A. & Harris, W.J. (1999). *Escherichia coli* Skp chaperone coexpression improves solubility and phage display of single-chain antibody fragments. *Protein Express. Purif.* **15**, 336-343.
46. Sachdev, D. & Chirgwin, J.M. (1998). Solubility of proteins isolated from inclusion bodies is enhanced by fusion to maltose-binding protein or thioredoxin. *Protein Express. Purif.* **12**, 122-132.
47. Davis, G.D., Elisee, C., Newham, D.M. & Harrison, R.G. (1999). New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnol. Bioeng.* **65**, 382-388.
48. Harrison, R.G. (2000). Expression of soluble heterologous proteins via fusion with NusA protein. *inNOVAtions* **11**, 4-7.
49. Bowden, G.A., Paredes, A.M. & Georgiou, G. (1991). Structure and morphology of inclusion bodies in *Escherichia coli*. *BioTechnol.* **9**, 725-730.
50. Wulfing, C. & Pluckthun, A. (1994). Correctly folded T-cell receptor fragments in the periplasm of *Escherichia coli*. Influence of folding catalysts. *J. Mol. Biol.* **242**, 655-669.
51. Wulfing, C. & Pluckthun, A. (1994). Protein folding in the periplasm of *Escherichia coli*. *Mol. Microbiol.* **12**, 685-692.
52. Bardwell, J.C.A. (1994). Building bridges: disulfide bond formation in the cell. *Mol. Microbiol.* **14**, 199-205.
53. Waldo, G., Standish, B.M., Berendzen, J. & Terwilliger, T.C. (1999). Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* **17**, 691-695.
54. Marston, F.A.O. & Hartley, D.L. (1990). Solubilization of protein aggregates. *Methods Enzymol.* **182**, 264-276.
55. Taylor, M.A.J., Pratt, K.A., Revell, D.F., Baker, K.C., Sumner, I.G. & Goodenough, P.W. (1992). Active papain renatured and processed from insoluble recombinant propapain expressed in *Escherichia coli*. *Protein Eng.* **5**, 455-459.
56. Zhi, W., Landry, S.J., Gierasch, L.M. & Srere, P.A. (1992). Renaturation of citrate synthase: influence of denaturant and folding assistants. *Protein Sci.* **1**, 522-529.
57. Kurucz, I., Titus, J.A., Jost, C.R. & Segal, D.M. (1995). Correct disulfide pairing and efficient refolding of detergent-solubilized single-chain Fv proteins from bacterial inclusion bodies. *Mol. Immunol.* **32**, 1443-1452.
58. Burgess, R.R. (1996). Purification of overproduced *Escherichia coli* RNA polymerase σ factors by solubilizing inclusion bodies and refolding from Sarkosyl. *Methods Enzymol.* **273**, 145-149.
59. Mukhopadhyay, A. (1997). Inclusion bodies and purification of proteins in biologically active forms. In *Advances in Biochemical Engineering and Biotechnology.* (Scheper, T., ed.), pp. 61-109, Springer-Verlag, Berlin, Germany.
60. De Bernardez Clark, E. (1998). Refolding of recombinant proteins. *Curr. Opin. Biotechnol.* **9**, 157-163.
61. Lilie, H., Schwarz, E. & Rudolph, R. (1998). Advances in refolding of proteins produced in *E. coli*. *Curr. Opin. Biotechnol.* **9**, 497-501.
62. Anderson, M., *et al.*, & Taylor, I. (1999). Refolding, purification, and characterization of a loop deletion mutant of human Bcl-2 from bacterial inclusion bodies. *Protein Express. Purif.* **15**, 162-170.
63. Geng, X. & Chang, X. (1992). High-performance hydrophobic interaction chromatography as a tool for protein refolding. *J. Chromatogr.* **599**, 185-194.
64. Armstrong, N., DeLencastre, A. & Gouaux, E. (1999). A new protein folding screen: application to the ligand binding domains of a glutamate and kainate receptor and to lysozyme and carbonic anhydrase. *Protein Sci.* **8**, 1475-1483.
65. Lepre, C. & Moore, J.M. (1998). Microdrop screening: a rapid method to optimize solvent conditions for NMR spectroscopy of proteins. *J. Biomol. NMR* **12**, 493-499.
66. Ferre-D'Amare, A.R. & Burley, S.K. (1994). Use of dynamic light scattering to assess crystallizability of macromolecules and macromolecular assemblies. *Structure* **2**, 357-359.
67. Giege, R. & Ducruix, A. (1999). An introduction to the crystallogenesis of biological macromolecules. In *Crystallization of Nucleic Acids and Proteins. A Practical Approach.* (Ducruix, A. & Giege, R., eds), pp. 1-16, Oxford University Press, Oxford, UK.
68. Lorber, B. & Giege, R. (1992). Preparation and handling of biological macromolecules for crystallization. In *Crystallization of Nucleic Acids and Proteins. A Practical Approach*, (Ducruix, A. & Giege, R., eds), pp 19-45, IRL Press, New York, USA.
69. McPherson, A. (1982). *Preparation and Analysis of Protein Crystals.* John Wiley and Sons, New York, USA.
70. Yates, J.R., III (2000). Mass spectrometry from genomics to proteomics. *Trends Genet.* **16**, 5-8.
71. Gevaert, K. & Vandekerckhove, J. (2000). Protein identification methods in proteomics. *Electrophoresis* **21**, 1145-1154.
72. Bergfors, T.M. (1999). Dynamic light scattering. In *Protein Crystallization: Techniques, Strategies and Tips.* (Bergfors, T.M. ed.), International University Line, La Jolla, CA.
73. Ferre-D'Amare, A.R. & Burley, S.K. (1997). Dynamic light scattering in evaluating crystallizability of macromolecules. *Methods Enzymol.* **276**, 157-166.
74. D'Arcy, A. (1994). Crystallizing proteins – a rational approach? *Acta Crystallogr. D* **50**, 469-471.
75. Bussow, K., Nordhoff, E., Lubbert, C., Lehrach, H. & Walter, G. (2000). A human cDNA Library for high-throughput protein expression screening. *Genomics* **65**, 1-8.
76. Watt, A.P., Morrison, D., Locker, K.L. & Evans, D.C. (2000). Higher throughput bioanalysis by automation of a protein precipitation assay using a 96-well format with detection by LC-MS/MS. *Anal. Chem.* **72**, 979-984.
77. Allan, G.F., Hutchins, A. & Clancy, J. (1999). An ultrahigh-throughput screening assay for estrogen receptor ligands. *Anal. Biochem.* **275**, 243-247.
78. Harry, J.L., Wilkins, M.R., Herbert, B.R., Packer, N.H., Gooley, A.A. & Williams, K.L. (2000). Proteomics: capacity versus utility. *Electrophoresis* **21**, 1071-1081.

79. Abramson, J. & Iwata, S. (1999). Crystallization of membrane proteins. In *Protein Crystallization: Techniques, Strategies and Tips.* (Bergfors, T.M. ed.), International University Line, La Jolla, CA.
80. Shatzman, A.R., Gross, M.S. & Rosenberg, M. (1997). Expression using vectors with phage λ regulatory sequences. In *Current Protocols in Molecular Biology*. pp. 16.3.8-16.3.11, John Wiley and Sons, Inc, New York, NY.
81. Crowe, J., Döbeli, H., Gentz, R., Hochuli, E., Stüber, D. & Henco, K. (1994). 6xHis-Ni-NTA chromatography as a superior technique in recombinant protein expression/purification. In *Methods in Molecular Biology*. (Harwood, A.J., ed.). Vol. 31, pp. 371-387, Humana Press, Inc., Totawa.
82. Sherwood, R. (1991). Protein fusions: bioseparations and application. *Trends Biotechnol.* **9**, 1-3.
83. Hopp, T.P., *et al.*, & Conlon, P.J. (1988). A short polypeptide marker sequence useful for recombinant protein identification and purification. *Bio/Technology* **6**, 1204-1210.
84. Lu, Z., *et al.*, & McCoy, J.M. (1996). Histidine patch thioredoxins. *J. Biol. Chem.* **271**, 5059-5065.
85. Schmidt, T.G.M. & Skerra, A. (1993). The random peptide library-assisted engineering of a C-terminal affinity peptide, useful for the detection and purification of a functional Ig Fv fragment. *Protein Eng.* **6**, 109-122.
86. Nilsson, B., Abrahmsen, L. & Uhlen, M. (1985). Immobilization and purification of enzymes with staphylococcal protein A gene fusion vectors. *EMBO J.* **4**, 1075-1080.
87. Greenwood, J.M., Gilkes, N.R., Kilburn, D.G., Miller, R.C., Jr. & Warren, R.A.J. (1989). Fusion to an endoglucanase allows alkaline phosphatase to bind to cellulose. *FEBS Lett.* **244**, 127-131.
88. Ong, E., Gilkes, N.R., Warren, R.A.J., Miller, R.C., Jr. & Kilburn, D.G. (1989). Enzyme immobilization using the cellulose-binding domain of a *Cellulomonas fimi* exoglucanase. *BioTechnol.* **7**, 604-607.
89. Greenwood, J.M., Ong, E., Gilkes, N.R., Warren, R.A.J., Miller, R.C., Jr. & Kilburn, D.G. (1992). Cellulose-binding domains: potential for purification of complex proteins. *Protein Eng.* **5**, 361-365.
90. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. & Prasher, D.C. (1994). Green fluorescent protein as a marker for gene expression. *Science* **263**, 802-805.
91. Ghrayeb, J., Kimura, H., Takahara, M., Hsiung, H., Masui, Y. & Inouye, M. (1984). Secretion cloning vectors in *Escherichia coli*. *EMBO J.* **3**, 2437-2442.