

Cloning, Production, and Purification of Proteins for a Medium-Scale Structural Genomics Project

Sophie Quevillon-Cheruel, Bruno Collinet, Lionel Trésaugues, Philippe Minard, Gilles Henckes, Robert Aufrère, Karine Blondeau, Cong-Zhao Zhou, Dominique Liger, Nabila Bettache, Anne Poupon, Ilham Aboulfath, Nicolas Leulliot, Joël Janin, and Herman van Tilbeurgh

Summary

The South-Paris Yeast Structural Genomics Pilot Project (<http://www.genomics.eu.org>) aims at systematically expressing, purifying, and determining the three-dimensional structures of *Saccharomyces cerevisiae* proteins. We have already cloned 240 yeast open reading frames in the *Escherichia coli* pET system. Eighty-two percent of the targets can be expressed in *E. coli*, and 61% yield soluble protein. We have currently purified 58 proteins. Twelve X-ray structures have been solved, six are in progress, and six other proteins gave crystals. In this chapter, we present the general experimental flowchart applied for this project. One of the main difficulties encountered in this pilot project was the low solubility of a great number of target proteins. We have developed parallel strategies to recover these proteins from inclusion bodies, including refolding, coexpression with chaperones, and an in vitro expression system. A limited proteolysis protocol, developed to localize flexible regions in proteins that could hinder crystallization, is also described.

Key Words: Yeast proteins; protein expression; structural genomics; inclusion bodies; co-expression.

1. Introduction

Structural genomics aims at the systematic structure determination of proteins, driven either by structural and/or functional objectives (*1*). The principal goals of the South-Paris Yeast Pilot Project are to express, purify, and systematically determine the structure of soluble single-domain proteins of the yeast *Saccharomyces cerevisiae* (<http://www.genomics.eu.org> [*2*]). At the present stage 240 yeast open reading frames (ORFs) have been cloned using a standard

protocol in a unique expression system, with constructs containing a hexahistidine (His₆)-tag at the 3' end of the target genes. In a single-pass experiment, 82% of these could be expressed in *Escherichia coli*, and 61% were soluble. We have currently purified 58 proteins. Twelve X-ray structures have been solved, six are in progress, and six additional protein crystals are being optimized. The resolution of two structures by nuclear magnetic resonance (NMR) is in progress.

One of the main tasks of the project is to set up efficient strategies for (1) the cloning of yeast ORFs, (2) the overexpression in *E. coli* of the corresponding recombinant proteins, and (3) their purification for structural studies. We have adopted a systematic approach that allows us to compare the efficiency of cloning and purification strategies on a large ensemble of proteins, all are prepared using the same protocol. Although cloning and expression were, in general, met with success, the low solubility of a large number of target proteins caused a considerable drop in the overall efficiency of the process that goes from a gene clone to a protein structure. We have, therefore, developed parallel strategies to recover proteins from inclusion bodies, including in vitro refolding or coexpression with chaperones in addition to in vitro expression techniques. A second predicted bottleneck is the low crystallization success rate for otherwise well-behaved and soluble proteins. We also describe here a simple limited proteolysis protocol, which localizes flexible parts of proteins and can be used to design shorter constructs that are more likely to crystallize.

2. Materials

1. Purified genomic DNA from yeast S288C, used as starting material for PCR and ORF cloning. This is the strain used for the *S. cerevisiae* genome sequencing project (3).
2. Genomic DNA purification buffer: 100 μ L of 50 mM Tris-HCl pH 8.0, 20 mM EDTA and 0.6% sodium dodecyl sulfate (SDS).
3. Oligonucleotide primers (MWG Biotech, Roissy CDG, France).
4. DyNAzyme (Finnzymes [Ozyme], St. Quentin en Yvelines, France).
5. DNA modification enzymes (Taq DNA polymerase, restriction enzymes, T4 DNA ligase) (New England Biolabs [Ozyme]).
6. Agarose gel equipment.
7. Expression vectors pET (Stratagene [Ozyme]).
8. Vector pCRT7/CT-TOPO (Invitrogen, Cergy Pontoise, France).
9. Plasmid pGKJE3 for overexpression of chaperones (4).
10. *E. coli* strains: XL1-Blue, BL21(DE3)pLysS, Rosetta(DE3)pLysS (Stratagene), C41(DE3), and C43(DE3) (Avidis, St. Beauzire, France).
11. Transformation buffer: 50 mM CaCl₂, at 4°C.
12. 2xYT (BIO101) and M63-derived minimum media (VWR-Prolabo, Fontenay-sous-Bois, France).
13. LeMaster amino acid mix (5) (final concentration in milligrams per liter): L-Ala 250, L-Arg 290, L-Asp 200, L-Gly 270, L-Cys 17, L-Pro 50, L-Ser 1080, L-Tyr 84, L-His 30, L-Gln 170, and L-Glu 330 (Sigma, St. Quentin Fallavier, France).
14. L-Selenomethionine (Se-Met) (Acros Organics, Noisy le Grand, France).

15. ^{15}N ammonium chloride and ^{13}C glycerol (Martek Biosciences Corporation, Columbia, MD).
16. Kanamycin, chloramphenicol, tetracyclin (Sigma).
17. IPTG (isopropyl- β -D-thio-galactopyranoside) (Sigma).
18. Bacterial incubators (Multitron, Infors, Massy, France) and bioreactors (1.5-L capacity; Applikon System, Les Mureaux, France).
19. Rodhorsyl (VWR-Prolabo).
20. Sonicator.
21. Liquid nitrogen.
22. Ni-nitroloacetic (NTA) resin (Qiagen, Courtaboeuf, France).
23. Lysis buffer: 20 mM Tris-HCl, pH 7.5, 200 mM NaCl, and 5 mM β -mercapto-ethanol (β -SH), at 4°C.
24. Wash buffer: lysis buffer supplemented with 20 mM imidazole, at 4°C.
25. Elution buffer: lysis buffer supplemented with 200 mM imidazole, at 4°C.
26. Inclusion bodies resolubilizing buffer: 6 M guanidinium hydrochloride (GndCl), 50 mM Tris-HCl, pH 7.5, 150 mM NaCl, and 5 mM β -SH, at 4°C.
27. In vitro refolding buffers: 200 mM NaCl or 20% glycerol or 0.6 M arginine or “cocktail buffer” (a mix of 50 mM each of CuSO_4 , ZnCl_2 , MgCl_2 , MnCl_2 , ADP, NADH, biotin, and thiamine), each at pH 6.5, 7.0, and 8.5.
28. SDS-polyacrylamide gel (PAGE) equipment.
29. Superdex 75 and 200 (16/60) (Amersham Biosciences, Orsay, France), chromatography equipment.
30. Vivaspin 6 and 20 concentrators (Sartorius, Palaiseau, France).
31. Proteases (Sigma).

3. Methods

The methods described next follow the chronological steps of our experimental flowchart, shown in **Fig. 1** and comprise: (1) the selection of 250 ORFs from the S288C *S. cerevisiae* genome, (2) their cloning in an *E. coli* expression vector, (3) the strategy for rapid testing of their overexpression and solubility, including the comparison of the efficiency of expression strains, (4) the multiple option strategy developed for recovery of inclusion bodies, (5) large-scale production of recombinant proteins, including (6) optimization of synthetic culture medium and labeling of proteins for X-ray or NMR, (7) purification and characterization of proteins, and, finally, (8) the development of a simple and rapid limited proteolysis protocol to localize nonstructured regions within purified proteins.

3.1. Target Selection

Not all proteins are equally well suited for a high-throughput structure determination approach. In order to test technologies and to establish protocols in our structural genomics project, a subset of yeast proteins was selected (2). Membrane proteins (detected by TMpred [6]) and multiple-domain proteins were excluded, as well as proteins containing low-complexity regions and coiled coil domains. ORFs were classified into three categories

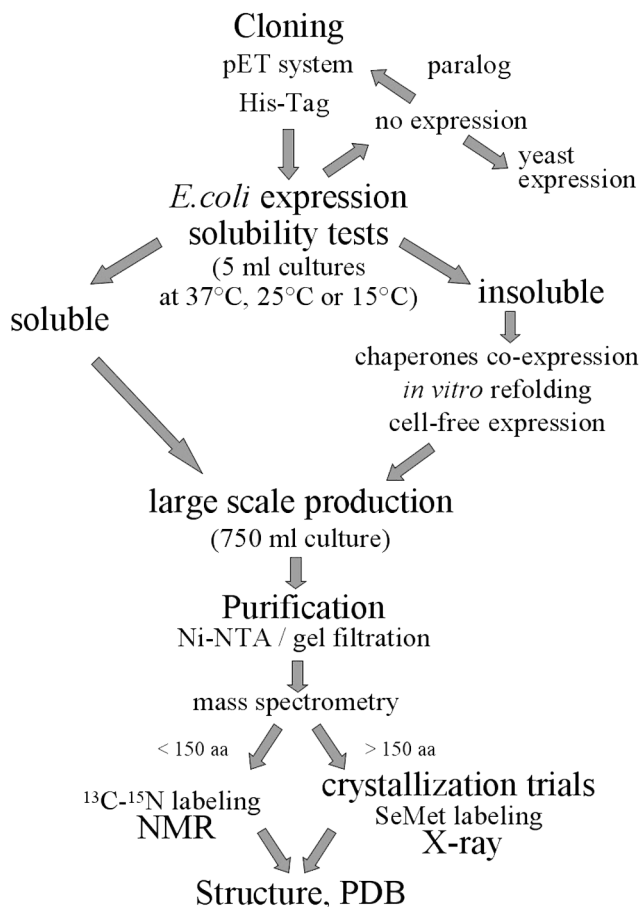


Fig. 1. Simplified experimental flowchart adopted within our Yeast Structural Genomics Project (<http://genomics.eu.org>).

by homology search using sequence comparisons (DARWIN [7], FASTA [8], and BLAST [9]): (1) those that are homologous to a protein of known structure, (2) those that are homologous to proteins whose structure is unknown, and (3) those that do not have a clearly identified homolog. This bioinformatics filter, combined with motif scans such as PRODOM (10), allows the presence of multiple-domain proteins to be detected. The third filter used a motif search algorithm (ProfileScan, PFAM [11]). These tools are actively used by our group to identify domains within large proteins for structural studies. The fourth filter is the search for homologies using multiple and/or iterative alignments (HMMER, PFAM, PSI-BLAST [12]), which are more sensitive than pairwise sequence alignments. The last filter used fold recog-

nition techniques (3DPSSM [13] and FROST [14]), which can be more powerful than standard sequence comparison methods alone, when sequence homology falls below detection level.

The list was further trimmed for the presence of transmembrane segments or the presence of a low-complexity segment (using Hydrophobic Cluster Analysis [15]), “sticky” proteins (with coiled coil regions for instance), or proteins that were already targeted by other structural genomics project.

3.2. Cloning

A general cloning strategy for a large-scale structural genomics project, based on a first step of PCR amplification from yeast genomic DNA, has to satisfy different criteria including (1) the choice of the prokaryotic expression system (T7 promoter), (2) the cloning strategy (classical restriction/ligation in pET vectors [16] or ligase free cloning method, e.g., “Topo-TA cloning” [17]), (3) the presence or not of an additional copy of the *lacI* gene on the vector, limiting the production of proteins before IPTG induction, and (4) the nature and position of a tag (at the N- or C-terminus of the protein). One also has to decide whether the tag will be cleaved or not after affinity purification (**Fig. 2**). Some thoughts and considerations around these strategic options are gathered in **Notes 1** and **2**, and some characteristics of the four vectors used in this study are listed in **Table 1**.

1. The genomic DNA is purified from a 0.5-mL overnight yeast culture (the cells are resuspended in genomic DNA purification buffer, incubated at 100°C during 10 min, and centrifuged for 10 min at 14,000g. The supernatant is diluted 10-fold in water) and is used as a template for PCR reactions (35 cycles, “hot-start” protocol).
2. The selected ORFs were inserted between a 5′-oligonucleotide containing a *NdeI* site in place of the AUG codon and a 3′-oligonucleotide. This sequence is immediately followed by six histidine codons, a stop codon, and, finally, a *NotI* sequence. The PCR reaction mixture, 50 µL in total, is composed of 0.5 µL of genomic DNA, 0.5 U of DyNAzyme EXT (Finnzymes), 30 pmol of each primer, and 0.01 mM dNTP in the suitable enzyme buffer.
3. The PCR products are purified with the PCR Purification Kit from Qiagen. The digestion with restriction enzymes is performed overnight at 37°C. The inserts are ligated after a second step of purification in a derivative pET-9 or pET-29 vectors (**Table 1** and **ref. 18**). When a *NdeI* site already existed in the selected ORF, the cloning is made in a pET-28 vector between *NcoI* and *NotI* sites.
4. The standard DNA manipulations are made in XL1-Blue strain (Stratagene).
5. The plasmids are purified with the Plasmid Purification Kit from Qiagen.
6. The DNA sequence of the constructs is checked.

3.3. Protein Production

The overexpression of the *S. cerevisiae* proteins in *E. coli* should be tested first on a small-scale culture (5 mL), in order to select (1) the best expression

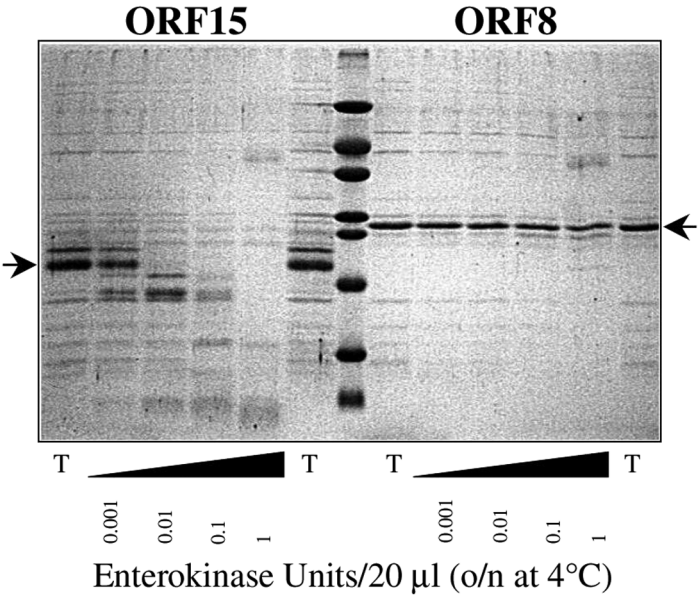


Fig. 2. Sodium dodecyl sulfate-polyacrylamide gel analysis of His-tag cleavage by enterokinase digestion on partially purified proteins containing the cleavage site DDDDK between the His-tag and the protein target. Two µg protein was incubated overnight at 4°C in 20-µL reaction mixtures with increasing amounts of protease. The figure illustrates the case of a protein completely degraded by proteolytic treatment (ORF15), and another that gave poor digestion yields (ORF8). T, lane without protease. The arrows point to nondigested proteins.

Table 1
Characteristics of the Plasmids Used in This Study

	Company	Promoter	Cloning sites	<i>lacI</i> gene	Resistance	Commercial tag
Derived pET-9 ^a	Novagen	T7 prom	<i>NdeI/NotI</i>	no	Kan	not used
pET-29	Novagen	T7 prom	<i>NdeI/NotI</i>	yes	Kan	not used
pET-28	Novagen	T7 prom	<i>NcoI/NotI</i>	yes	Kan	not used
pCR ^R T7/ CT-TOPO	Invitrogen	T7 prom	TA-cloning	no	Amp	not used

^aModified polylinker *NdeI* – *SfiI* – *NotI*.

strain and (2) the optimal temperature of induction for soluble expression. We have routinely used the following two strains: BL21(DE3)pLysS and Rosetta(DE3)pLysS. Rosetta coexpresses tRNAs corresponding to codons

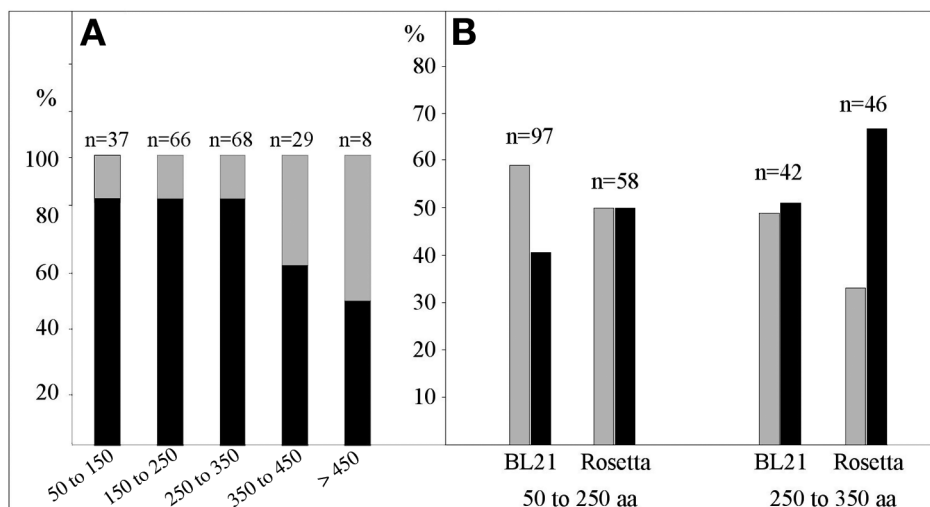


Fig. 3. Small-scale expression tests. **(A)** Percentage of expressed (black) vs not expressed (gray) proteins for targets of different sizes (number of amino acids). **(B)** Comparison of the expression efficiency for yeast proteins in two different *Escherichia coli* strains: BL21(DE3)pLysS and Rosetta(DE3)pLysS. Gray, low and medium expression levels; black, high expression levels. The targets are divided into two groups: constructs made of 50–250 amino acids, and those composed of 250–350 amino acids.

rarely used in *E. coli* but frequently used in eukaryotes. Some trials using the C41(DE3) and C43(DE3) strains, originally developed for the expression of membrane or toxic proteins (19), were carried out for proteins not expressed by the aforementioned system, but were not met with success. Nevertheless 80% of protein targets smaller than 350 amino acids were successfully expressed. This percentage drops to 50% for larger proteins (Fig. 3A). An interesting observation resides in the comparison of expression rates in the BL21 and Rosetta strains as a function of the length of the protein targets. Even if some bias exists (the tests were not always performed on the same set of proteins) it is clear that the presence of rare tRNAs in the Rosetta strain favored a higher expression of larger proteins compared to BL21 (see Fig. 3B and Note 3).

3.3.1. Transformation of *E. coli* Expression Strains and Protein Induction

1. Competent cells are prepared with standard transformation protocols: heat-shock (cold CaCl_2) or electroporation protocol. It is important to note that the use of freshly transformed cells is mandatory for obtaining an efficient and reproducible expression. Previously transformed expression strains that were kept as glycerol stocks are not reliable as starting expression material. It seems that the expression strain cells transformed with pET vectors are not stable during extended storage at -80°C .

2. From a culture of untransformed expression strain, a stock of competent cells is prepared, aliquoted in 500 μL , and stored at -80°C . For expression, an aliquot of competent cells is thawed and 50 μL are transformed with about 100 ng of each pure plasmid. The transformed cells are not plated but directly grown overnight at 37°C in 5 mL liquid 2YT medium supplemented with Kan or Amp. This is subsequently used as a preculture.
3. The following day 10 mL of medium are inoculated with 250 μL of preculture and cells grown until $A_{600\text{nm}}$ is reached at 1. Protein expression is induced with 0.3 mM IPTG. The culture is then divided into two 5-mL aliquots, and each are incubated at 37 or 25°C . Protein expression is allowed to take place for 4 h (or alternatively overnight when a lower expression temperature is chosen).
4. The cultures are centrifuged at 5000g during 10 min at 4°C . The cells are resuspended in 1 mL of lysis buffer and stored at -20°C overnight. This freezing step will help the subsequent lysis step.

3.3.2. Screen for Protein Expression Level and Solubility

1. The suspended cells are thawed at room temperature and sonicated for 15–30 s at 4°C . The solution then becomes less viscous presumably because of the breakage of *E. coli* genomic DNA.
2. An aliquot of the “total extract” is analyzed by SDS-PAGE according to standard protocol.
3. The rest of the lysed cells are centrifuged at 13,000g, 4°C , during 30 min. An aliquot of the clear supernatant (“soluble extract”) is analyzed by SDS-PAGE.
4. 14% Acrylamide gels are loaded with 5 or 10 μL of samples and are stained with Brilliant Blue (Sigma). An example of the expression of two ORFs is shown **Fig. 4**, ORF3 is expressed in a soluble form, and ORF4 is expressed as inclusion bodies.

3.4. Strategies for Recovery of Inclusion Bodies

Because 37% of the 204 expressed proteins form inclusion bodies in *E. coli*, we developed a procedure for the recovery of these proteins. The strategy has been described in detail elsewhere (**18**), and only a brief overview will be given here. A set of 20 representative proteins expressed as inclusion bodies was studied in parallel. The strategy is made of three different options, adapted from refolding protocols for structural genomics (high-throughput) purposes. (1) In vitro refolding by dilution: after purification of the inclusion bodies in denaturing conditions (6 M GdnCl), the proteins are refolded by dilution using a screen of refolding buffers following a procedure adapted to a 96-well plate format; folding is followed by measuring light scattering at 390 nm (a high absorbance is an indication of the formation of protein aggregates); (2) coexpression of the target protein with bacterial chaperones (DnaK-DnaJ-GrpE-GroEL-GroES), using the plasmid developed by Nishihara et al. (**4**); and (3) cell-free expression, which is a different way to express the proteins. We chose to use the technology developed in Dr. Yokoyama’s laboratory (**20,21**),

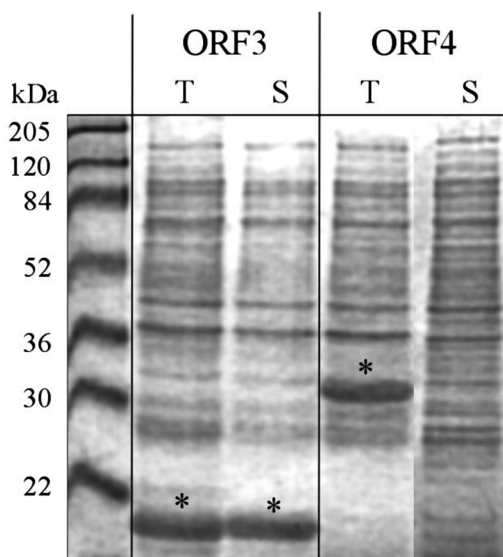


Fig. 4. Sodium dodecyl sulfate-polyacrylamide gel of crude extracts of *Escherichia coli* showing an example of a soluble protein (ORF3) and an example of expression in inclusion bodies (ORF4). T, total cell extracts obtained after freezing–thawing and sonication; S, soluble extracts after centrifugation for 30 min at 13,000g of total extracts.

in a batch scale (50- μ L reaction volume) or dialysis scale (1–2 mL reaction volume), producing, respectively, micrograms or milligrams of protein. According to the ORF under study, the three approaches were useful for recovering inclusion bodies, and complemented each other. Some proteins were rescued by all three protocols, whereas others were refolded by only one or two of them. The chaperones' coexpression approach is easily adaptable to a pre-existent expression protocol and, therefore, is particularly useful for high-throughput structural genomics. To complete this short overview, *see* **Note 4** discussing some refolding strategies developed in other structural genomics projects, especially those that use the expression of fusion proteins (22). Other important strategies consist in directed evolution (23) or in switching to eukaryotic expression systems (24).

3.5. Large-Scale Production

Basically, 750 mL of 2xYT medium are inoculated in flasks at 37°C with 10 mL of freshly transformed overnight precultures. The target proteins are expressed after IPTG induction for 4 h, the optimal temperature determined during expression and solubility screen. This procedure typically yields between 5 and 50 mg of recombinant proteins. When overexpression is too low or when a large scale of

protein is needed for crystallization screens or biochemical studies, bioreactor facilities are used. The pH in the reactor is maintained at 7.0 by adding either NaOH or H₂SO₄. The dissolved oxygen content is maintained greater than 30% air saturation by increasing the agitation speed from 800 to 1500g. Aeration is kept to 1 v.v.m. (1 vol of air per 1 vol of culture per minute). Foaming is controlled by addition of one-tenth diluted Rodhorsyl (VWR-Prolabo). In flasks or bioreactors, the growth process consists in two steps, the biomass production achieved at 37°C, and the protein production performed at optimal temperature (*see Subheading 3.3*). The induction period (2 h, 4 h, or overnight) is dependent on the incubation temperature (37, 25, or 15°C, respectively). Synthetic media should be optimized for scaling up as described in **ref. 25**.

3.6. Labeling

3.6.1. Se-Met Labeling for Crystallography Studies

Multiwavelength anomalous diffraction (MAD) phasing using selenomethionine-substituted protein crystals is the method of choice for the determination of X-ray structures (**5**). Two alternative strategies are frequently used for the incorporation of Se-Met into proteins. The use of *E. coli* B834 strain (Novagen), which is auxotrophic for methionine (**26**), was not satisfactory in our hands because we obtained low growth rates, owing to the toxicity of Se-Met and poor protein yields. The method we finally adapted in our project relies on the metabolic inhibition of the methionine pathway to obtain Se-Met incorporation using a standard expression strain (**27,28**).

1. 500 mL of M63mGly5 culture (this medium derived from M63 medium described in **refs. 29** and **30** and supplemented with the LeMaster amino acid solution known to activate the general cell metabolism (*see Subheading 2*. and **ref. 5**) of *E. coli* expression strain transformed with the pET construct is grown at 37°C.
2. At OD₆₀₀=1 to 1.5, the suspension is supplemented by a cocktail of amino acids (L-Lys, L-Phe, and L-Thr at 125 mg/L each; L-Ile, L-Leu, L-Val at 62.5 mg/L each), to repress the methionine biosynthesis pathway. L-Se-Met is added at 62.5 mg/L.
3. The production of the recombinant protein is induced 30 min later by addition of 0.3 mM IPTG, for 2 h at 37°C, 4–6 h at 25°C, or overnight at 15°C.
4. Se-Met incorporation into the protein is assayed by mass spectrometry (MS) after purification of the protein.

3.6.2. ¹³C and ¹⁵N Labeling for NMR Studies

The resolution of a structure by NMR requires uniform labeling of the protein nitrogens (¹⁵N) and carbons (¹³C). Culture media volumes are kept to a minimum (250–500 mL), mainly because of the cost of the labeled carbon source (¹³C-glycerol). In order to choose when to induce, the consumption of labeled glycerol is followed by HPLC during bacterial growth.

1. The first overnight inoculum is cultivated in 10 mL of 2xYT medium at 37°C and 200 g.
2. An overnight preseed culture in the appropriate medium (50 mL of M63m¹⁵NGly5 or 20 mL of M63m¹⁵N/¹³C-Gly5) is inoculated at an initial OD₆₀₀ of 0.1.
3. The totality of the preseed culture is added into the final culture composed of the same medium. The cells are grown at 37°C until the exponential growth phase.
4. At an OD₆₀₀ of about two, the temperature is eventually reduced prior to the addition of 0.3 mM IPTG. The induction is maintained 2 h to overnight depending on the temperature.
5. The labeling is assayed by MS after purification of the protein.

3.7. Protein Purification and Biophysical Controls

The high-throughput nature of a structural genomics project demands a general and simple purification protocol (*see Note 5; 31–34*). We, therefore, chose to add a His₆-tag to the recombinant proteins. The first purification step is a Ni⁺⁺ affinity column and the tag is generally not removed for the subsequent crystallization experiments (*see Note 6 and Fig. 2*). From a few test experiments to cleave the His-tag by proteolytic digestion, we concluded that it would be very difficult to integrate this step into a systematic and rapid protocol. We speculated that the crystallization of the majority of proteins will not be affected by the presence of the short tag. The affinity step is systematically followed by a gel filtration chromatography step to remove contaminant proteins and aggregates and to estimate the monodispersity and oligomeric state of the proteins. In most cases, this protocol yielded sufficient quantities of purified protein.

1. Cells obtained from a 750-mL culture are stored at –20°C at least overnight in 40 mL of lysis buffer, and broken by three cycles of freezing/thawing and sonication at 4°C. The suspension is centrifuged at 13,000g for 30 min at 4°C.
2. The supernatant is loaded on 2 mL of Ni-NTA equilibrated in the lysis buffer. The flow-through is kept on ice for SDS-PAGE control. The resin is washed with 20 mL of the buffer supplemented with 20 mM imidazole. The protein is eluted in three steps with 8 mL of the buffer containing respectively 100, 200, and 400 mM imidazole. An aliquot of each fraction is loaded on a SDS-PAGE to localize the protein.
3. The protein-containing fraction(s) are concentrated by centrifugation with a Vivaspinn concentrator (Vivascience). The protein is immediately applied (*see Note 6*) to a Superdex 75 or 200 and is eluted at 1 mL/min for each protein in a suitable buffer in terms of pH (an electrofocusing analysis may be necessary) and NaCl concentration (*see Note 6*).
4. A SDS-PAGE of the fractions containing the protein is performed in order to control the purity and to correctly choose the fractions to be pooled.
5. The pure protein is concentrated for crystallization trials (usually around 10 mg/mL). Crystallization trays are set up as quickly as possible after protein sample preparation; best results for crystallization are obtained with very fresh protein samples.

6. An aliquot of pure protein is systematically assayed by MS, in order to control the integrity of the sample and/or the correct incorporation of various labels (Se-Met or $^{13}\text{C}/^{15}\text{N}$). In some ambiguous cases, we carry out one-dimensional or two-dimensional NMR spectra to control the correct folding of the proteins. Other biophysical data (circular dichroism, microcalorimetry, isothermal calorimetry, small angle X-ray scattering, or fluorescence) sometimes complement our standard analysis protocol.

3.8. Limited Proteolysis

Many well-structured proteins contain regions of high conformational mobility, often situated at the N- or C-terminus of the protein. It is well established that these regions often hinder crystallization. This fact that we are not removing the intrinsically mobile terminal His-tag might actually make things worse (as previously mentioned, we found that omitting the proteolysis step considerably speeds up the purification process and also results in higher purification yields). Although we successfully crystallized 37% of the 59 purified proteins, we wanted to test biochemical protocols to increase the yield and quality of protein crystals. We developed a simple and small-scale limited proteolysis protocol to generate large subfragments of the proteins, which are resistant to further proteolysis and may, therefore, correspond to the structured and globular protein cores (**Fig. 5**). The partial cleavage is first measured via SDS-PAGE and if proteolysis has taken place, a binding test onto the Ni-NTA column helps to localize the proteolytic cleavage site. Afterward, precise localization of the digested site is carried out by MS. The fragment is subcloned by PCR and the new construct expressed in *E. coli* for large-scale expression and purification. At this time, four proteins that failed to crystallize have been subcloned. The polypeptides remained well overexpressed and are purified in the same buffers as the natives one. The crystallization trials are in progress.

1. In a 50- μL mixture, 10 μg of pure protein are incubated with 1/10 and 1/200 (w/w) of protease (trypsin, papain, pepsin, and so on), for 30 min at 37°C.
2. 10 μL is immediately analyzed on a SDS-PAGE gel.
3. 5 μL is frozen for analysis by MS.
4. The rest is bound to 20 μL of Ni-NTA, the resin is washed, and the polypeptide is eluted with 400 mM imidazole. All the fractions are analyzed by SDS-PAGE (**Fig. 5**).

4. Notes

1. The standard expression system we used is based on the pET system. In addition to the high level T7 promoter common to the pET series, the important features of our standard construct are: (1) a Kan R marker more adaptable to high cell density fermentation than Amp R marker, (2) the expressed protein is strictly limited to the ORF sequence fused to the His-tag, without any linkers that might inhibit crystallization. We decided to keep the tag in place because proteolytic procedures cannot be systematically applied to a large number of targets, as illustrated for two

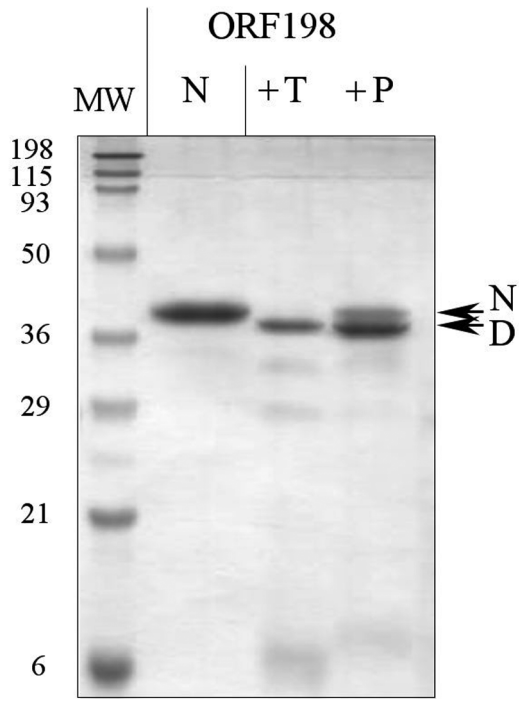


Fig. 5. Sodium dodecyl sulfate-polyacrylamide gel showing the limited proteolysis digestion of purified ORF198. Lane N corresponds to the native protein before incubation. Lane +T, digestion by trypsin; lane +P, digestion by papain. Arrows point to the native protein (N) and to the globular core of the protein identified by limited proteolysis (D).

ORFs in **Fig. 2**. The His-tag was introduced just after the last amino acid of the protein (and not at the N-terminus), in order to retrieve only full-length transcripts during the Ni-NTA purification step. A test experiment on 30 proteins for which we compared the expression and solubility yields between constructs with the His-tag at the N- and C-terminus, respectively, did not allow us to discern any marked differences. For ORF PCR and subsequent cloning, we used 3' primers (50mers) made of the last six codons, six histidine codons, a stop codon, the *NotI* restriction site, and four extra bases (the 5' primers was shorter with just *NdeI* or *NcoI* site, ATG and some coding codons). This strategy requires ordering of long primers (about 50 nt), more prone to sequence errors during chemical synthesis.

2. Even if the restriction/ligation cloning in pET vectors was very efficient (for each construction, four clones were tested and three or four contained the insert), a facilitated and less time-consuming cloning strategy, based on the "Topo-TA cloning" technology (pCRT7/CT-TOPO vector), was tested at the beginning of the project on 59/81 selected ORFs. Surprisingly, we have observed a difference of efficiency for expression of proteins for the ORFs cloned in this TOPO vector when compared

with those cloned in a pET vector: 82% of proteins were expressed in the “pET system,” whereas only 52% were expressed in the “Topo system.” To confirm these observations we subcloned 36 ORFs in the pET vector, which did not express in “Topo vectors,” or only expressed at a very low level. Half of the Topo-unexpressed proteins and all of the 10 low-expressed proteins became highly expressed in the pET vector, whereas 13 proteins remained unexpressed.

3. The systematic approaches applied in structural genomics projects around the world on the production of large numbers of proteins, allow for the first time to compare the efficiency of protocols classically used in laboratories for the production of recombinant proteins. In our project, we focused on the comparison of expression efficiency of commercial *E. coli* strains (see **Subheading 3.3.** and **Fig. 3**). At this stage we can also provide some concluding remarks concerning the “pET expression system” in general. (1) We initiated the project with a systematic comparison of expression level between systems using plasmids containing the *lacI* gene or not (pET-9 vs pET-29). Because the strict criteria of the target selection led us to a list of *a priori* cytoplasmic proteins (see **Subheading 3.1.**), the presence of a supplementary copy of the *lacI* gene on the vectors was not crucial, and gave in general a slightly lower expression level of the proteins. (2) We systematically verified expression leakage of the recombinant proteins during the growth phase of the cultures before addition of IPTG, a frequent problem (observed in about 50% of the cases) constituting a drawback for the production of Se-Met-labeled proteins. This convinced us to adapt the labeling protocol by growing the cultures as soon as possible in minimum medium complemented with Se-Met in place of Met, at the very beginning of the exponential phase (see **Subheading 3.6.**).
4. The large number of proteins expressed as inclusion bodies in *E. coli* is one of the most important bottlenecks we were confronted with. The most simple experimental parameter to influence solubility of expressed proteins is to lower the induction temperature. On a set of 140 well-expressed proteins we observed (without discriminating between the strains) that 48% of the proteins were soluble when produced at 37°C, and interestingly 22% became soluble when the expression temperature was lowered to 25°C or below. For the remaining 30%, we developed the three-layered strategy as described in **Subheading 3.4.**, and finally decided to routinely coexpress the five chaperones. For instance, the presence of chaperones increased the solubility, between 10 and 90%, for 17/29 insoluble proteins. Alternatively, other structural genomics projects use fusion proteins with soluble domains (green fluorescence protein, maltose-binding protein, glutathione-S-transferase, and other) allowing targets to be kept in the soluble phase (22) (see Chapter 1). For this type of strategy, which requires making several constructs for each target, the “Gateway” cloning technology (Invitrogen) combined with automation of the procedure is recommended. The most important problem is the necessity to release the fused domain before crystallization trials. Even if the crystallization of several proteins fused to maltose-binding protein via a rigid and short spacer has recently been described, this will generally not be the case (35).
5. The results presented herein were part of the Yeast Structural Genomics Pilot Project that took place during years 2001 and 2003. Since then, several other studies have

started (see <http://genomics.eu.org/spip/-Projects->) and the protocols described in this chapter are still in routine use for these new structural genomics projects.

6. At this stage of the project, 60% of the 121 proteins expressed under soluble form have been tested for purification in order to obtain milligram quantities of pure protein for crystallization trials. The first affinity purification step is exactly the same for all proteins (see **Subheading 3.7.**). The only difficulty consists in the determination of the optimal pH and salt concentration of the gel filtration buffer, consistent with a mono-disperse protein, in each individual case. Eighty-two percent of 72 proteins were purified to homogeneity and at sufficient quantities for setting up automated crystallization screens (at least 200 μ L at 2–50 mg/mL). The one-dimensional or two-dimensional NMR spectra obtained for some proteins allowed detection of the existence of very soluble, highly concentrated but “unfolded” proteins (36). We developed a biophysical-based study including small angle X-ray scattering, microcalorimetry, or circular dichroism to better understand these phenomena and to verify if any ligands, cofactors, or nucleic/protein partners are necessary for the protein to adopt a well-defined structure.

Acknowledgments

This work is supported by grants from the Ministère de la Recherche et de la Technologie (Programme Génopoles).

References

1. Brenner, S. E. (2001) A tour of structural genomics. *Nat. Rev. Genet.* **2**, 801–809.
2. Quevillon-Cheruel, S., Collinet, B., Zhou, C. Z., et al. (2003) A structural genomics initiative on yeast proteins. *J. Synchrotron Radiat.* **10**, 4–8.
3. Goffeau, A., Barrell, B.G., Bussey, H., et al. (1996) Life with 6000 genes. *Science* **274**, 563–567.
4. Nishihara, K., Kitagawa, M., Yanagi, H., and Yura, T. (1998) Chaperone coexpression plasmids: differential and synergistic roles of DnaK-DnaJ-GrpE and GroEL-GroES in assisting folding of an allergen of Japanese cedar pollen, Cryj2, in *Escherichia coli*. *Appl. Environ. Microbiol.* **64**, 1694–1699.
5. Hendrickson, W. A., Horton, J. R., and LeMaster, D. M. (1990) Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J.* **9**, 1665–1672.
6. Hofmann, K. and Stoffel, W. (1993) Tmbase: A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler* **374**, 166.
7. Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992) Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443–1445.
8. Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
9. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
10. Rost B. and Liu J. (2003) The PredictProtein server. *Nucleic Acids Res.* **31**, 3300–3304.

11. Bateman A., Coin L., Durbin R., et al. (2004) The Pfam protein families database. *Nucleic Acids Res.* **32**, 138–141.
12. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
13. Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. E. (2000) Enhanced Genome Annotation using Structural Profiles in the Program 3D-PSSM. *J. Mol. Biol.* **299**, 499–520.
14. Marin, A., Pothier, J., Zimmerman, K., and Gibrat, J. F. (2001) Protein structure prediction: bioinformatic approach. In: *Protein Threading Statistics: An Attempt to Assess the Significance of a Fold*, (Tsigelny, I., ed.), International University Line, La Jolla, CA.
15. Callebaut, I., Labesse, G., Durand, P., et al. (1997) Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell. Mol. Life Sci.* **53**, 621–645.
16. Studier, F. W. and 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.
17. Shuman, S. (1994) Novel approach to molecular cloning and polynucleotide synthesis using vaccinia DNA topoisomerase. *J. Biol. Chem.* **269**, 32,678–32,684.
18. Trésaugues, L., Collinet, B., Minard, P., et al. (2004) Refolding strategies from inclusion bodies in a structural genomics project. *J. Struct. Funct. Genomics* **5**, 195–204.
19. Miroux B. and Walker J. (1996) Over-production of proteins in *Escherichia coli*: mutants hosts that allow synthesis of some membrane proteins and globular proteins at high levels. *J. Mol. Biol.* **260**, 289–298.
20. Kigawa, T., Muto, Y., and Yokoyama, S. (1995) Cell-free synthesis and amino acid-selective stable isotope labeling of proteins for NMR analysis. *J. Biomol. NMR* **6**, 129–134.
21. Kigawa, T., Yabuki, T., Yoshida, Y., et al. (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. *FEBS Lett.* **442**, 15–19.
22. Hammarstrom, M., Hellgren, N., van Den Berg, S., Berglund, H., and Hard, T. (2002) Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Sci.* **11**, 313–321.
23. Waldo, G. S. (2003) Improving protein folding efficiency by directed evolution using the GFP folding reporter. *Methods Mol. Biol.* **230**, 343–359.
24. Boettner, M., Prinz, B., Holz, C., Stahl, U., and Lang, C. (2002) High-throughput screening for expression of heterologous proteins in the yeast *Pichia pastoris*. *J. Biotechnol.* **99**, 51–62.
25. Bettache, N. A., Quevillon-Cheruel, S., Bondet, V., van Tilbeurgh, H., and Blondeau, K. Determination of optimal cultivation conditions for large scale production of yeast carboxyl methyltransferase (Ppm1) in *Escherichia coli*., submitted.
26. Studts, J. M. and Fox, B. G. (1999) Application of fed-batch fermentation to the preparation of isotopically labeled or selenomethionyl-labeled proteins. *Protein Expr. Purif.* **16**, 109–119.

27. Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L., and Clardy, J. (1993) Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J. Mol. Biol.* **229**, 105–124.
28. Doublié, S. (1997) Preparation of selenomethionyl proteins for phase determination. *Meth. Enzymol.* **276**, 523–530.
29. Sambrook, J., Frits, E. F., and Maniatis, T. (eds.) (1989) Preparation and transformation of competent *E.coli*. In: *Molecular Cloning, 2nd ed.*, CSH Laboratory Press, Cold Spring Harbor, NY, pp. I.74–I.84.
30. David, G., Blondeau, K., Renouard, M., and Lewit-Bentley, A. (2002) Crystallization and preliminary analysis of *Escherichia coli* YodA. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 1243–1245.
31. Lesley, S. A., Kuhn, P., Godzik, A., et al. (2002) Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. *Proc. Natl. Acad. Sci. USA* **99**, 11,664–11,669.
32. Vincentelli, R., Bignon, C., Gruez, A., et al. (2003) Medium-scale structural genomics: strategies for protein expression and crystallization. *Acc. Chem. Res.* **36**, 165–172.
33. Heinemann, U., Bussow, K., Mueller, U., and Umbach, P. (2003) Facilities and methods for the high-throughput crystal structural analysis of human proteins. *Acc. Chem. Res.* **36**, 157–163.
34. Matte, A., Sivaraman, J., Ekiel, I., Gehring, K., Jia, Z., and Cygler, M. (2003) Contribution of structural genomics to understanding the biology of *Escherichia coli*. *J. Bacteriol.* **185**, 3994–4002.
35. Smyth, D. R., Mrozkiewicz, M. K., McGrath, W. J., Listwan, P., and Kobe, B. (2003) Crystal structures of fusion proteins with large-affinity tags. *Protein Sci.* **12**, 1313–1322.
36. Uversky, V. N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756.

Macromolecular Crystallography Protocols, Volume 1

Preparation and Crystallization of Macromolecules

Doublie, S. (Ed.)

2007, XIV, 280 p. 49 illus., Hardcover

ISBN: 978-1-58829-292-6

A product of Humana Press