

Letter to the Editor

## **Streamlining the production of proteins for structural biology**

Raymond J Owens<sup>1,2\*</sup> & Opher Gileadi<sup>3</sup>

*<sup>1</sup>Research Complex at Harwell, Rutherford Appleton Laboratory, OX11 0FA, UK and*

*<sup>2</sup>Division of Structural Biology, Henry Wellcome Building for Genomic Medicine,  
University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, UK*

*<sup>3</sup>Structural Genomic Consortium, University of Oxford, Old Road Campus Research  
Building, Roosevelt Drive, Oxford OX3 7DQ, UK*

\*Corresponding author

Email address: ray@strubi.ox.ac.uk

There has been a transformation in the power and throughput of analytic methods delivered by large-scale facilities, such as synchrotrons. This has placed an increasing demand on the supply of high quality samples (purified proteins and protein crystals) for structural studies. To meet this requirement, protein production has been streamlined to improve efficiency and throughput. Nonetheless, delivering recombinant proteins in sufficient quantity and quality remains a significant bottleneck in the gene to structure pipeline for a variety of reasons including low levels of expression, poor solubility and toxicity when over-expressed. Many of these issues had been encountered by participants on the START workshop on Biophysics & Structural Biology at Synchrotrons, during the course of their projects. Therefore, our aim over the first day of the workshop was to share our experience of setting up and running protein production workflows that address the challenge of producing a wide variety of proteins for structural analysis. Lessons learnt were explained and some examples of how to clone, express and purify different proteins were discussed with the participants in some “virtual” practical work and case studies.

A typical protein production workflow can be broadly divided into three stages (1) design (2) screening expression and (3) sample preparation (Figure 1). Within each stage there are two essential operations, beginning with the design of expression vectors. For this step, prior knowledge from related proteins in the Protein Data Bank and bioinformatics analyses using machine learning to predict structural features in a sequence, are combined to define the gene segment that will be cloned and expressed. Many of the projects of the students involved making a relatively small number of vectors and expressing them in just one host, typically, *E. coli*. However, by designing and constructing multiple vectors in parallel and expressing them in more than one host (e.g. insect and *E. coli*), the optimum sequence and expression host for producing a target protein can be more quickly identified. Suites of plasmid vectors have been developed that are adapted for ligation-independent cloning and enable expression testing in *E. coli*, mammalian and insect cells (Nettlehip et al 2019). Inclusion of a short histidine tag comprising 6-10 residues in all expression vectors provides a facile way of detecting and purifying the expressed protein.

Expression screening of multiple constructs is carried out at small-scale in either 24 or 96 well plate formats reducing experimental costs by ensuring that subsequent scale-up and purification is only carried using optimised construct/host cell combinations.

*E. coli* has been used successfully to produce a large variety of recombinant proteins, and remains the host of choice for prokaryotic and some eukaryotic proteins, particularly structural domains (Gräslund et al 2008). By using insect and mammalian cell systems as well, expression of eukaryotic proteins that fail to be produced in *E. coli* can often be rescued (Savitsky et al 2010).

Given that automation and/or the use of multichannel pipettes can be used to carry out the cloning and screening steps efficiently, this stage of the process is not rate-limiting. Whereas, the scale-up and purification of selected proteins still requires some form of batch culture of cells typically in shake flasks and column-based chromatography steps, to isolate the protein sample, all of which are relatively low throughput operations. However, having pre-screened for expression, a simple two stage purification strategy combining an affinity capture step, typically immobilised metal affinity chromatography (IMAC) and size exclusion chromatography is often sufficient to produce samples of > 95 % purity (by SDS-PAGE). Structural biology techniques generally require the purified protein sample to be monodisperse and a size fractionation step during purification provides the first indication of protein behaviour. Analysis of the purified by intact protein mass spectrometry can then be used to authenticate the sample and identify any post-translation modifications that change the expected mass of the protein. It is a truism that the protein is the key variable and therefore quality assessment steps are necessary to both characterise the sample and to minimise batch to batch variation.

In summary, the main take home points from our session were as follows:-

- Construct design is critical and should be informed by all available information from the literature and bioinformatic analyses.
- Protein production throughput can be increased by using standard operating procedures and parallel processing of vector construction and expression screening.
- Incorporating quality assessment of sample purity and homogeneity into the workflow is essential to characterise the protein and ensure consistency of batch to batch preparation.

## References

Gräslund S, Nordlund P, Weigelt J et al. (2008) Protein production and purification. Nat Methods. 5: 135-46.

Nettleship JE, Rada H, Owens RJ (2019) Overview of a high throughput pipeline for streamlining the production of recombinant proteins. *Methods in Molecular Biology*, vol. 2025 (in press).

Savitsky P1, Bray J, Cooper CD, Marsden BD, Mahajan P, Burgess-Brown NA, Gileadi (2010) High-throughput production of human proteins for crystallization: the SGC experience. *J Struct Biol* 172: 3-13.

### **Figure legend**

Overview of the protein production pipeline (from Nettleship et al 2019 reproduced with permission from Springer Protocols)