Journal of
Synchrotron
Radiation

ISSN 0909-0495

Received 18 February 2004
Accepted 22 October 2004

Towards the high-throughput expression of metalloproteins from the *Mycobacterium tuberculosis* genome

John F. Hall,[a]* Mark J. Ellis,[b] Takanori Kigawa,[c] Takashi Yabuki,[c] Takayoshi Matsuda,[c] Eiko Seki,[c] S. Samar Hasnain[b] and Shigeyuki Yokoyama[c,d,e]

[a]Cell Signalling Laboratory, De Montfort University, Leicester LE1 9BH, UK, [b]Molecular Biophysics Group, Daresbury Laboratory, Warrington WA4 4AD, UK, [c]RIKEN Genomic Sciences Center, Tsurumi, Yokohama 230-0045, Japan, [d]RIKEN Harima Institute at SPring-8, Mikazuki-cho, Sayo, Hyogo 679-5148, Japan, and [e]The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. E-mail: jfhall@dmu.ac.uk

The provision of high-quality protein in adequate quantities is a prerequisite for any structural genomics programme. A number of proteins from the *Mycobacterium tuberculosis* genome have been expressed and the success at each stage of the process assessed. Major difficulties have been encountered in the purification and solubilization of many of these proteins, most likely as a result of mis-folding. Some improvements have been made to the protocol but the overall success rate is still limited; however, the use of a cell-free protein expression system will circumvent some of the difficulties encountered. Alternative purification systems are also required and the properties of a mutant blue copper protein are described, that may offer a combined purification and tagging system.

Keywords: *Mycobacterium tuberculosis*; metalloproteins; cell-free protein expression; purification tags.

1. Introduction

The provision of a reliable supply of high-quality protein in substantial quantities is a pre-requisite of any structural programme. The major difficulties encountered for such a high-throughput protein expression system are in the initial stages, where the gene(s) of interest are cloned and the recombinant protein expressed from suitable microbial hosts such as *E. coli* and *P. pastoris*, and the final purification stage, which ideally consists of a single simple procedure that results in the production of clean protein. The introduction of target genes into expression vectors, of which the pET (Novagen) or Gateway (Invitrogen) series are perhaps the most widely used and reliable, is a time-consuming process which limits the productivity of any high-throughput system. The use of cells as the expression system introduces additional extraction steps and furthermore cannot easily handle proteins that may be toxic or have a tendency to aggregate. The introduction of metals, either those naturally occurring in metalloproteins or in the form of selenomethionine for MAD phasing, also presents difficulties.

The ligation-independent system offered by Roche Diagnostics goes some way to addressing these problems. However, the synthesis of large quantities of protein for structural studies is still not possible from linear templates using this system and substantial improvements were necessary to enable the direct and rapid production of protein from linear templates generated by the polymerase chain reaction (PCR). These problems have been addressed at RIKEN and substantial improvements have been made, so that it is possible to obtain 6 mg ml^{-1} of protein labelled with stable isotope for NMR studies (Kigawa *et al.*, 1999) and introduce the cytotoxic selenomethionine into protein with high efficiency (>95%; Kigawa *et al.*, 2002). The levels of protein produced are comparable with those obtained from other systems, removing one of the primary objections to the use of cell-free protein expression (CFPE) systems, and further improvements to the *E. coli* S30 preparations used will realise greater yields (Kigawa *et al.*, 2004). A further advantage of the CFPE system is the flexibility and ease with which alternative vectors can be produced. Thus, in cases where solubility may pose a problem, solubility enhancement tags can be added, along with other purification tags and folding reporters such as green fluorescent protein (GFP) *etc.* (Yokoyama, 2003). Although difficulties remain, notably with integral membrane proteins and multi-domain proteins, the CFPE system currently offers the best and most rapid means for high-throughput screening projects.

In the case of metalloproteins, an additional advantage of the CFPE is that metal incorporation can be explored much more widely without the complication of toxic effects on the expression organism. The incorporation of metal centres in

Table 1
Revised protocol for the amplification of selected genes from MTB.

Initial PCR parameters	Revised PCR parameters
(1) 371 K for 5 min	(1) 368 K for 1 min
(2) 369 K for 25 s	(2) 323 K for 30 s
(3) 345 K for 25 s	(3) 338 K for 30 s
(4) 345 K for 2 min	(4) 368 K for 30 s
(5) Go to (2) ×7 reduce by 2 K each cycle	(5) 323 K for 30 s
(6) 369 K for 25 s	(6) 338 K for 30 s
(7) 329 K for 25 s	(7) Go to (4) ×40
(8) 345 K for 2 min	
(9) Go to (6) ×25	

expressed proteins remains a concern in any high-throughput structural programme as the lower organisms used for high-throughput expression quite often do not possess the specific chaperones that may be needed by the metalloprotein for metal incorporation. This paper presents an overview of the results so far using the established procedures, and some preliminary results from the CFPE.

2. Results

The technology used relies upon the now well established pET expression vectors and *E. coli* BL21(DE3). Briefly, the protocol is as follows. The targeted sequences are amplified by PCR utilizing primers to incorporate Nde I and BamHI restriction sites (or others if these cannot be used). The PCR products are then purified and selected according to size on agarose gels and restricted with the appropriate enzyme prior to ligation into a modified pET28 vector. This vector, in addition to encoding for kanamycin resistance and His tags, also carries a gene for spectinomycin resistance and a GFP tag which can be activated to form a fusion protein with the protein from the target gene in the presence of tetracycline. These constructs are then used to transform *E. coli* BL21(DE3) directly. Subsequently the fluorescing colonies are selected automatically and cultured to produce the target protein.

The initial attempts to express protein, using the existing protocol, resulted in successful production from less than 20% of our selected targets. We therefore undertook to express the remainder of the targets and determine the possible causes of the unacceptably high attrition rate. The standard touchdown PCR procedure gave inconclusive results in our hands. Many of the reactions gave multiple bands or in some cases nothing at all. Scrutiny of the conditions and repetition of the PCR gave the same results which were clearly unsatisfactory. We therefore revised the PCR protocol and determined that a 'one size fits all' procedure seemed inappropriate. The best results were obtained with a relatively simple protocol (Table 1) which, with adjustments to the annealing temperature and time, gave reliable results for most targets (Fig. 1). Following purification, the products were ligated into pGEM-T and sequenced prior to further manipulation. The results of the sequencing showed that the majority of the sequences were as expected. The few discrepancies were probably the result of misincorporation by *Taq* polymerase which tends to have the highest error rate of the available thermostable polymerases. However, in one case the sequence had a 97% identity with a cloning vector indicating that selection on the basis of size alone may not be totally reliable. Nevertheless, at this stage there was a 90% success rate (Table 2). The genes were then excised from the pGEM-T vectors with the appropriate enzymes and ligated into the standard pET28a expression vector, transformed into *E. coli* BL21(DE3) and protein expression induced with IPTG according to standard procedures. The majority of proteins were expressed at good to moderate levels as is evident in lanes 1–6 and 9 and 10 of Fig. 2. However, there are no additional bands in lanes 7 and 8 corresponding to overexpressed protein. This phenomenon has been noted before with other constructs that appear to be complete, in frame and identical to other clones that do produce protein.

Table 2
Comparison of the current and previous success rates, expressed as a percentage of the initial number of target genes, illustrating the minor improvements made after modifying the procedures at each stage.

	Current	Previous
PCR	95	95
Size	90	90
Expression	76	76
Solubility	38	19
Purification	14	5

The degree of success at this stage is running at some 76% (Table 2) of the original targets, which is entirely consistent with the general trend which ranges from 59 to 94%. The use of the CFPE system is likely to improve this still further, as in trials with particularly intransigent clones, which had failed to produce any protein using the conventional expression methods, there was an immediate and substantial improvement in all cases but one (Fig. 3). The extraction and purification phases, however, present serious difficulties. Problems were encountered with solubility in several cases and, of those that were soluble, the one-step His 6 tag/nickel affinity column process did not always give protein of sufficient quality for further analysis and crystallization (Fig. 4). Solubility and the

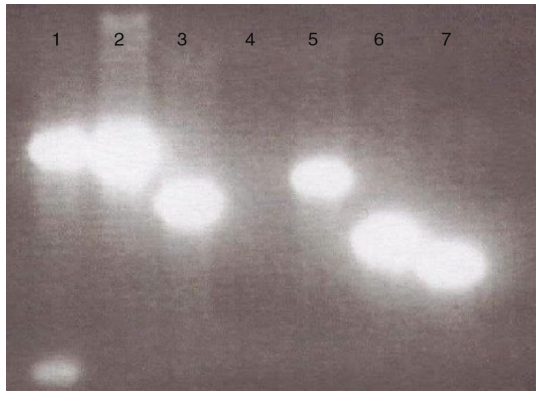

Figure 1
Revised PCR protocol. Results of the revised protocol demonstrating the improvement from the multiple bands obtained with the original protocol (Table 1). 1: rv2229; 2: rv1388; 3: rv2305; 4: rv1371; 5: rv2986; 6: rv3628; 7: standard.

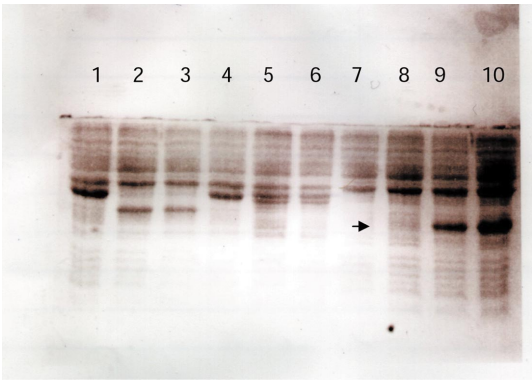

Figure 2
Demonstration of the variability in the level of protein expressed between clones. 1: rv2229c clone 1; 2: rv1388 clone 1; 3: rv1388 clone 2; 4: rv2229c clone 2; 5: rv2986c clone 1; 6: rv2986c clone 2; 7–10: rv3628 clones 1–4. Note in particular the absence of any significant expression in clone 2 of rv3628 compared with clones 3 and 4.

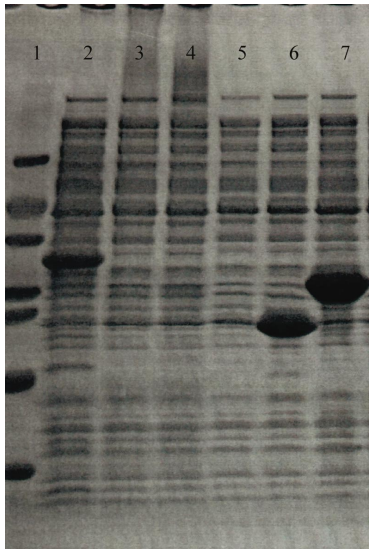

Figure 3
The expression of *Mycobacterium tuberculosis* (MTB) proteins using a linear DNA template and the CFPE system (RIKEN). 1: Molecular weight markers; 2, 3, 4 and 6: MTB proteins; 5: SOD; 6: GFP.

inextricably linked folding problems are well documented and, in the case of *M. tuberculosis*, may present a particular problem. There are, however, a very large number of diverse strategies to ameliorate the situation. These include the use of various solubility enhancement tags such as domain B1 and maltose binding protein (MBP) (Routzahn & Waugh, 2002), and fluorescent tags such as GFP that can increase the yield of protein and act as folding reporters (Kaba *et al.*, 2002). Misfolded or truncated proteins are usually deposited in inclusion bodies and refolding from these can be achieved using chaotropic reagents followed by dilution. Attempts to improve upon the process have included column methods (Middelberg, 2002), solid-state chaperones (Park *et al.*, 2002), high pressure (Randolph *et al.*, 2002) and so on. All of these methods have drawbacks, principally in terms of the need for further manipulation and extra relatively sophisticated and costly equipment, thus reducing the efficiency and high-

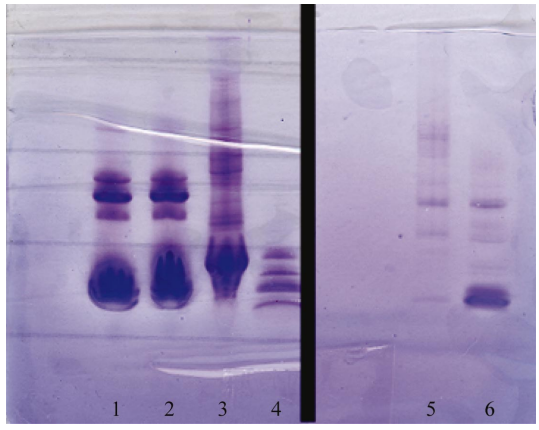

Figure 4
Purification of MTB proteins. 1: Protein from rv2547(His 6 tagged) after nickel affinity column purification; 2: protein from rv2547 after a second pass down a nickel affinity column; 3: protein from rv38365 after nickel affinity column purification; 4: protein molecular weight markers; 5: protein from rv3836 after nickel affinity and gel filtration; 6: protein from rv2547 after nickel affinity and gel filtration.

throughput potential of the system. Furthermore, should the initial tag system not deliver pure protein there is often no easy means of following the progress of the subsequent purification using standard methods. Additional tags have been proposed; for example, it has been recognized that MBP is more efficacious if a second tag is used (GST, His) and coloured tags such as rubredoxin have been promoted (Kohli & Ostermeier, 2003). An ideal situation would be to have a tag that identifies the fusion protein and has its own intrinsic purification properties.

We have been working with the small blue copper protein rusticyanin to determine the structural features that are responsible for the extraordinarily high redox potential and extreme pH tolerance. It has been shown (Grossmann *et al.*, 2002) that the so far unique N-terminus of the protein (in comparison with other cupredoxins) shields the hydrophobic core and is in fact responsible for the stability of the protein in alkaline environments. Thus, removal of the complete 35 amino acid extension results in a protein that is soluble at pH values below 4.5 and will precipitate at pH values of 5.0 and above. Since then, we have constructed a −7 and −20 variant. These also show reduced alkaline stability and the −20 form will also precipitate from solution at about pH 5.6. Furthermore, the −20 mutant retains some of its colour and can be resolubilized in more acidic buffer. This mutation of rusticyanin, or variants thereof, may provide us with a protein tag that will aid purification by a means that is markedly different in principle to other systems currently in use. We are currently constructing a vector that will demonstrate whether the system will function as a fusion protein.

3. Conclusions

A number of variations on the protocol outlined above have been applied to the task of providing sufficient amounts of quality protein for structural studies from *M. tuberculosis*. In

each case the overall statistics are comparable, with typical success rates of: cloning 95%, expression 80%, solubility 22% and crystallization 5%, all of which are in line with our findings for the putative metalloproteins (Table 2). We have demonstrated that fine tuning of these procedures can enhance the success rate to some degree and that the application of the CFPE system (RIKEN) will increase this still further. Nevertheless, difficulties remain, notably with the solubilization and purification of the expressed proteins, and repurification is often essential. The combination of several techniques has continued to bring success and will undoubtedly remain the optimal approach in the near future. However, there is a clear need for purification systems that operate *via* alternative principles combined with a readily detectable tag should further purification be needed. It is interesting to speculate that the mutant rusticyanin protein could offer one such possibility and may introduce another valuable tool with which to address the purification problems faced in all of the structural genomics programmes.

This work was supported by the RIKEN Structural Genomics/Proteomics Initiative (RSGI), the National Project on Protein Structural and Functional Analyses, Ministry of Education, Culture, Sports, Science and Technology of Japan and was conducted as part of the CLRC/RIKEN collaborative agreement.

References

Grossmann, J. G., Hall, J. F., Kanbi, L. D. & Hasnain, S. S. (2002). *Biochemistry*, **41**, 3613–3619.

Kaba, S. A., Nene, V., Musoke, A. J., Vlak, J. M. & van Oers, M. M. (2002). *Parasitology*, **125**, 497–505.

Kigawa, T., Yabuki, T., Matsuda, N., Matsuda, T., Nakajima, R., Tanaka, A. & Yokoyama, S. (2004). *J. Struct. Funct. Genom.* **5**, 63–68.

Kigawa, T., Yabuki, T., Yoshida, Y., Tsutsui, M., Ito, Y., Shibata, T. & Yokoyama, S. (1999). *FEBS Lett.* **442**, 15–19.

Kigawa, T., Yamaguchi-Nunokawa, E., Kodama, K., Matsuda, T., Yabuki, T., Matsuda, N., Ishitani, R., Nureki, O. & Yokoyama, S. (2002). *J. Struct. Funct. Genom.* **2**, 29–35.

Kohli, B. M. & Ostermeier, C. (2003). *Protein Expr. Purif.* **28**, 362–367.

Middelberg, A. P. J. (2002). *Trends Biotechnol.* **20**, 437–443.

Park, S. J., Ryu, K., Suh, C. W., Chai, Y. G., Kwon, O. B., Park, S. K. & Lee, E. K. (2002). *Biotechnol. Bioprocess Eng.* **7**, 1–5.

Randolph, T. W., Carpenter, J. F. & St John, R. (2002). US Patent 6 489 450.

Routzahn, K. M. & Waugh, D. A. (2002). *J. Struct. Funct. Genom.* **2**, 83–92.

Yokoyama, S. (2003). *Curr. Opin. Chem. Biol.* **7**, 39–43.