

## High-throughput production of *Pyrococcus furiosus* proteins: considerations for metalloproteins

F. E. Jenney Jr,\* P. S. Brereton,‡ M. Izumi,¶ F. L. Poole II, C. Shah, F. J. Sugar, H.-S. Lee and M. W. W. Adams

Southeastern Collaboratory for Structural Genomics, Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602-7229, USA. E-mail: fjenney@bmb.uga.edu

Free-living prokaryotic organisms contain all of the proteins required for the basic biochemical processes of life. As part of the Southeastern Collaboratory for Structural Genomics (SECSG), *Pyrococcus furiosus* is being used as a model system for developing a high-throughput protein expression and purification protocol. Its 1.9 million basepair genome encodes ~2200 putative proteins, less than 25% of which show similarity to any structurally characterized protein in the Protein Data Bank. The overall goal of the structural genomics initiative is to determine, in total, all existing protein folds. The immediate objective of this work is to obtain recombinant forms of all *P. furiosus* proteins in their functional states for structural determination. Proteins successfully produced by overexpression in another organism such as the bacterium *Escherichia coli* typically contain a single subunit, are soluble and do not contain (complex) cofactors. Analyses of the *P. furiosus* genome suggest that perhaps only a quarter of the genes encode proteins that would fall into this category. The hypothesis is that lack of the appropriate cofactor or of the partner protein(s) necessary to form a complex are major reasons why many recombinant proteins are insoluble. This work describes development of the production pipeline with attention to prediction and incorporation of cofactors.

**Keywords:** *Pyrococcus furiosus*; protein solubility; structural genomics; overexpression; cofactors; metalloprotein.

### 1. Introduction

The rapidly increasing availability of complete genomic sequences from organisms in all three domains of life gives a wealth of information on the diversity of the suite of proteins available to living cells. It is fair to say that, at the current time, only approximately 25% of open reading frames (ORFs) in most genomes encode proteins for which a function can be recognized by sequence comparison, and there are 25% more where function may be indicated or implied by comparison. This means that 50% of the ORFs in most genomes are essentially completely unknown, both structurally and functionally. While structural determination of unknown proteins cannot definitively indicate *in vivo* function, it can provide tremendous insight into possible function(s) as well as identifying novel types of protein folding. Accordingly, a number of centers internationally (see, for example, Heinemann *et al.*, 2000) and in the USA, as part of the Protein Structure Initiative ([www.nigms.nih.gov/psi](http://www.nigms.nih.gov/psi)), have been created to develop cost-effective high-throughput (HTP) methodologies

for rapid cloning, overexpression, purification and structural determination of the entire proteome from a number of model organisms. The goal of this effort, dubbed 'structural genomics', is to accelerate techniques for structure determination at every step from protein production to structural data acquisition and analysis.

The Southeastern Collaboratory for Structural Genomics (SECSG) brings together groups at the University of Georgia, Georgia State University, the Universities of Alabama at Huntsville and Birmingham and Duke University with the goal of developing methodologies using proteins from a prokaryote model, *Pyrococcus furiosus*, and two eukaryote model organisms, *Caenorhabditis elegans* and *Homo sapiens* (Adams *et al.*, 2003). *P. furiosus*, the subject of this work, is a member of the domain Archaea, and is both a hyperthermophile (optimal growth at 373 K) and a strict anaerobe (Fiala & Stetter, 1986). As a free-living organism, *P. furiosus* contains all the genes necessary for life and is a well studied organism biochemically (Adams *et al.*, 2001; Verhagen *et al.*, 2001). More recently, its metabolism is being investigated by microarray analyses of gene expression representing the entire genome under a number of different growth conditions (Schut *et al.*, 2001, 2003). The specific goal of the *P. furiosus* protein

\* Current address: Monsanto, Chesterfield, MO, USA.

¶ Current address: Marine Science Institute, University of California, Santa Barbara, CA 93106, USA.

production group in the SECSG is to clone, express and purify functional recombinant forms of all the proteins of *P. furiosus*.

It is relatively easy to heterologously express and purify a homomeric protein in the most commonly used expression host, *Escherichia coli* (Baneyx, 1999; Cornelis, 2000; Jonasson *et al.*, 2002), if it is small, negatively charged, water soluble and contains no cofactors (the so-called 'low-hanging fruit' or LHF). Historically, if one wanted to overexpress a more complex protein, this required more extensive individualized manipulation, either by coexpression of the known partners, or known chaperone genes (see, for example, Henricksen *et al.*, 1994; Li *et al.*, 1997; Stevens *et al.*, 2003), though in a sense this could be said of all heterologous expression. While the structural genomics initiative is aimed at developing HTP protein production techniques, relatively little attention has been paid so far to the more complex proteins, those containing metal or organic cofactors, membrane proteins, proteins that are part of heteromeric complexes, and combinations thereof (dubbed the 'high-hanging fruit' or HHF). Considering metal cofactors alone, approximately one-third of all proteins so far structurally characterized contain a metal cofactor, and perhaps as many as half of all proteins could contain metal (Holm *et al.*, 1996; Degtyarenko, 2000). It is very likely that many proteins which fail to express, or express only as insoluble inclusion bodies, are part of this large class (HHF), failing to fold as they lack a necessary cofactor, or a partner protein to stabilize them. There are a number of examples where individual members of a protein complex were, individually, either poorly expressed, expressed as inclusion bodies or expressed but with poor function (see, for example, Henricksen *et al.*, 1994; Li *et al.*, 1997), and co-expression of these genes in the same *E. coli* cell resulted in significant increases in yield of soluble protein and in functionality. Other techniques have been used to increase solubility of recombinant proteins, including the use of fusion proteins (Fox *et al.*, 2003; Pedelacq *et al.*, 2002), and mutagenesis to remove surface hydrophobic residues which may cause aggregation (Daujoytè *et al.*, 2003), to give only a few examples. While these strategies are clearly successful, they are not likely to work in every case for proteins which are part of native complexes, or which require accessory genes for cofactor insertion. Thus it is our goal to develop universal techniques for expression of all proteins from any given organism using the *P. furiosus* proteome as a model. These proteins should be in a functional form, properly folded, containing cofactors, and, where appropriate, as part of heteromeric complexes. In this work we focus in particular on the problems specific to metalloprotein prediction and production.

## 2. Materials and methods

The complete *P. furiosus* genome was obtained from the NCBI GenBank file (RefSeq NC\_003413; Robb *et al.*, 2001) and the general strategy was to divide all 2182 ORFs [2065 from the Genbank annotation and 117 putative ORFs (data

not shown)] into 25 'projects' in 96-well plates, each containing approximately 94 genes. They were sorted first by internal restriction sites (to facilitate cloning) and second by length, as gene amplification by the polymerase chain reaction (PCR) was most successful when all ORFs on the plate were approximately the same length. Primer pairs representing 5' and 3' ends of every ORF in the *P. furiosus* genome were designed by simply taking the first 21 nucleotides after the start codon, and adding the sequence containing the appropriate restriction enzyme site (for example *Bam*HI) to the 5' end. The 3' primers were made by adding the sequence for a unique *Not*I restriction site to the last 24–26 nucleotides (including the stop codon) of the ORF. The *Escherichia coli* protein expression plasmid pET-24d (Novagen, Madison, WI, USA) was modified using standard molecular biology techniques (Sambrook & Russell, 2001) to create a series of fusion protein expression vectors such that a fusion tag of MAHHHHHHXX- was placed at the N-terminus of each cloned *Pyrococcus* protein. The XX represents three different amino acid additions to the vectors resulting in a unique restriction enzyme site after the polyhistidine tag (pET-24dBam encoding GlySer; pET-24dHind encoding LysLeu, and pET-24dEco encoding GluPhe). All genes were cloned using standard molecular biology techniques [PCR, restriction digestion, ligation, transformation, restriction analysis to screen for inserts (Sambrook & Russell, 2001)], simply performed in a 96-well format instead of using individual tubes. The histidine tag allows a simple purification by specific binding to the immobilized metal (nickel or cobalt) on a chromatography column, which can result in as high as 95% purity in one step. The polyhistidine tagged proteins were purified using this standard IMAC (immobilized metal affinity chromatography) with Co or Ni affinity media (Clontech, Palo Alto, CA, USA or Qiagen, Valencia, CA, USA) using the manufacturer's protocol, followed by size-exclusion chromatography as a second purification step. Purified proteins were sent for inductively coupled plasma (ICP) emission spectroscopy or ICP mass spectrometry for metal content analysis (Chemical Analysis Facility, University of Georgia) and analyzed for the presence of Fe, Zn, Co, Cu, Mn, Mg, Cr, Mo, Cd, W and Ni. Measurements of  $\geq 0.2$  metal atoms per protein monomer were considered positive.

## 3. Results

### 3.1. Prediction of the metalloproteome

We use this term to refer to the entire collection of metalloproteins in the *P. furiosus* genome [see Scott *et al.*, 2005 (this issue)]. There are few studies available on techniques for prediction of bioinorganic protein motifs which may bind metal cofactors (Degtyarenko, 2000). Two techniques have been used here for a preliminary analysis of the *P. furiosus* genome in order to predict metalloprotein candidates which may contain zinc or iron (other metals were not considered for this first prediction). First, a simple count of cysteine motifs was made, defined as CysX<sub>n</sub>Cys (where  $n = 0-4$

**Table 1**

Comparison of a prediction of putative metal-liganding motifs with a prediction of putative metalloproteins based on keyword search of the INTERPRO database (see text).

The 'Known' class represents those ORFs with homology to known proteins in the INTERPRO database (970 genes), and the 'Unknown' class those with no known homolog (1212 genes).

Total ORFs in the <i>P. furiosus</i> genome with Cys motifs	356/2182	16%
'Unknown' ORFs with Cys motifs	119/970	12%
'Known' ORFs with Cys motifs	237/1212	19%
'Known' ORFs INTERPRO predicted Fe	225/1212	19%
'Known' ORFs INTERPRO predicted Zn	119/1212	10%
'Known' ORFs with Cys motifs and INTERPRO predicted Fe	78/237	33%
'Known' ORFs with Cys motif and INTERPRO predicted Zn	46/237	19%
'Known' ORFs with INTERPRO predicted Fe and Cys motifs	78/225	35%
'Known' ORFs with INTERPRO predicted Zn and Cys motifs	46/119	39%
Total ORFs with His motifs	151/2182	7%

**Table 2**

Results to date for *P. furiosus* cloning and protein production.

Percent success out of the total 2182 ORFs is indicated in parentheses.

Targets	PCR	Clone	Expressed†	Purified
2182	2039 (93%)	1873 (86%)	654 (30%)	186 (9%)

† Unique ORFs out of 2182. There have been a total of 1153 growths for expression to date.

amino acids). Cysteines often act as ligands for metal binding (for example in rubredoxin; Holm *et al.*, 1996; Jenney & Adams, 2001; Giles *et al.*, 2003) or in zinc finger proteins (Krishna *et al.*, 2003). Histidine residues are similarly involved in the zinc finger motif (Krishna *et al.*, 2003) and a similar count of putative histidine motifs was performed (Table 1). Second, a simple search of the INTERPRO database (a resource integrating a number of protein databases at <http://www.ebi.ac.uk/interpro/>; Mulder *et al.*, 2003) using keywords such as 'iron', 'Fe', 'zinc', 'Zn' *etc.* was used to search for known metalloprotein homologs in the *P. furiosus* genome. This search is only useful for the 'known' half of the genome, as any hypothetical or conserved hypothetical proteins would not have such keywords. The results of these searches are shown in Table 1. They indicate that these predictions do not overlap very well. While the overlap of the class of proteins known from previous work to contain Fe or Zn (based on the INTERPRO keyword search) with the cysteine motif prediction is reasonably high (74%), only about half of the proteins which do have cysteine motifs have been shown to contain Fe or Zn, at least with this relatively simplistic prediction technique. One major problem with using putative metal-binding motifs for prediction is that they are quite variable, in which amino acids are used as ligands, in the distance between the ligands (CysX<sub>n</sub>Cys) and in the distance between such motifs. New domains are discovered often (Makarova *et al.*, 2002) and, particularly in zinc fingers, the motifs can use either cysteine or histidine so almost any permutation can be

**Table 3**

Comparison of prediction of putative metal-liganding motifs (Cys motifs) and putative metalloproteins (INTERPRO keyword search for Fe or Zn, see text) with preliminary results from purification of 186 recombinant *P. furiosus* proteins.

The number of proteins predicted to have Fe or Zn by the INTERPRO search, which overlap with predicted cysteine motifs, is indicated in the last two columns in parentheses.

	Total	Cys motifs	INTERPRO Fe (Cys motifs)	INTERPRO Zn (Cys motifs)
Prediction for 186 proteins	186	38	29 (7)	21 (8)
Actual Fe-containing proteins	17	6	4 (1)	5 (2)
Actual Zn-containing proteins	26	12	6 (2)	3 (2)
Both Fe and Zn	7	3	2	2 (1)

possible C/HX<sub>n</sub>C/H X<sub>n</sub> C/HX<sub>n</sub>C/H (Krishna *et al.*, 2003). In summary, prediction based on homology to known metalloproteins, or known motifs, can be indicative of candidate metalloproteins, but will not be complete.

### 3.2. Production of proteins

Remarkably, although no particular care was taken to optimize primers for annealing, success in amplifying groups of 94 genes, as defined by a single PCR product of the predicted molecular weight, was as high as 100% for some projects. To date, 93% of the targeted ORFs have been successfully amplified by PCR and 86% have been successfully cloned into the modified pET vectors as judged by restriction analysis of the clones (Table 2). Furthermore, expression of 654 unique His-tagged *P. furiosus* ORFs has been attempted in *E. coli*. Of these, 344 (53%) were successfully expressed as judged by denaturing polyacrylamide gel electrophoresis analysis of the Ni-affinity column eluate, the first step in purification. A significant number of these proteins, however, precipitate at some step after this first column, resulting in purification of only 186 of those expressed proteins (28% success, but with 61 still in progress).

### 3.3. Metal content of recombinant proteins

Table 3 contains the current results for metal content measurements of recombinant *P. furiosus* proteins produced in *E. coli*. In this set of 186 proteins purified so far, the predictions indicate that 38 of them have cysteine motifs, only some of which overlap with predictions by the INTERPRO keyword search. Of the proteins purified, 17 actually contained iron, and 26 zinc, and seven of these zinc-containing proteins also contained iron based on chemical analysis. Mixed metal isoforms are not uncommon in recombinant proteins (see, for example, Eidsness *et al.*, 1992; Czaja *et al.*, 1995). The data in Table 3 clearly show that as yet there is no strong correlation between the predictive techniques used here and the presence or identity of a metal cofactor. Another significant problem affecting the metal content of recombinant proteins is the purification technique used. Table 4 demon-

**Table 4**

Comparison of the metal content of purified recombinant proteins on cobalt *versus* nickel chromatography affinity media.

Protein	Affinity matrix	Metal content	Affinity matrix	Metal content
Hydrogenase subunit alpha	Cobalt	1.5 Co	Nickel	None
Asparaginase	Cobalt	0.7 Co	Nickel	None
Glucose-1-P-thymidyl-transferase	Cobalt	0.8 Co	Nickel	0.2 Zn
Myo-inositol-1-phosphate synthase	Cobalt	0.2 Co	Nickel	None
Alcohol dehydrogenase	Cobalt	1.2 Co, 0.5 Zn	Nickel	0.13 Ni, 0.12 Zn

strates that purification of a number of different proteins using a cobalt affinity matrix results in significant cobalt contamination of the proteins, where nickel affinity material gives relatively little nickel contamination (though whether this is due to less leaching of metal from the column or less nickel binding to the proteins is not yet known).

#### 4. Discussion

Traditionally, a particular protein would typically be targeted for heterologous overexpression in view of a known or suspected functional role based on *in vivo* or *in vitro* evidence. The structural genomics strategy of determining structures of a vast number of proteins introduces a number of complications for protein overexpression. Chief among these problems is target selection and prediction. In some cases, in order to optimize success, the ‘low-hanging fruit’, *i.e.* proteins predicted to be relatively small, soluble, without cofactors and not part of heteromeric protein complexes, are targeted, in search of novel folds. The problem implicit in such a selection is that proteins which may be of great interest, but which require cofactors or partner proteins, are passed over and, in any case, if targeted, may not properly express or fold. Such a strategy, however, presupposes the ability to correctly predict which proteins are membrane-bound, part of complexes or contain cofactors. This can be particularly difficult in the classes of conserved hypothetical proteins and hypothetical proteins, for which there is no functional data. Membrane proteins can be predicted to some extent with a number of programs based on possible transmembrane regions (Holden *et al.*, 2001). In prokaryotes, protein complexes can be postulated based on the close proximity of the genes encoding them in putative operons. Prediction of putative cofactors, of particular interest to this work, is not as simple, as shown above by the relatively poor correlation between predictions and the initial purification results.

There are a number of practical considerations for overexpression and purification of metalloproteins in *E. coli*. First, it has been shown that, at least in some cases, growth conditions can affect which metal is incorporated into a metal binding site. For example, with overexpression of the small iron protein rubredoxin, growth in an undefined medium results in a mixture of Fe- and Zn-containing rubredoxin (Eidsness *et al.*, 1992). However, this can be relieved in some

cases by growth in a defined medium supplemented with the appropriate metal (Eidsness *et al.*, 1992; Jenney & Adams, 2001). In any case, for more complex metal centers, which require chaperone proteins for assembly (for example, nitrogenase; Schmid *et al.*, 2002), *E. coli* may not have the requisite genes for assembly. A second consideration is sensitivity of metalloproteins to oxygen. Many iron-containing proteins may be oxygen-sensitive, and in fact utilize this sensitivity *in vivo* for signaling in both aerobic and anaerobic organisms (for example, Hantke, 2001; Kang *et al.*, 2003). Anaerobic purification of proteins is labor-intensive and not as amenable to high-throughput techniques as aerobic purification, and the metal-chelating material used for purification is sensitive to some typical reductants such as dithiothreitol. Choice of metal for the IMAC purification can affect results as well, as shown in Table 4. These problems need to be addressed in a high-throughput protocol and the current effort is aimed at optimizing conditions for metalloprotein expression and purification.

#### 4.1. Are cofactors necessary, at least for structural genomics projects?

Certainly, the answer is yes if the goal is structures of the functional form of proteins. Given that so many proteins in a genome (30–50%, see *Introduction*) may contain a metal cofactor, it is critical that proper assembly of metalloproteins be considered in high-throughput protocols. In many cases it has been demonstrated that cofactors stabilize the native protein and may be essential for folding (see, for example, Wittung-Stafshede, 2002; Liu & Xu, 2002), and may speed up folding *in vitro* (Apiyo & Wittung-Stafshede, 2002). Not all proteins which are purified, however, will crystallize, and those that crystallize will not always diffract at sufficient resolution to provide structural information (Yee *et al.*, 2003). Thus, in collaboration with another group at the University of Georgia [see Scott *et al.*, 2005 (this issue)], development of a high-throughput protocol for analyzing recombinant metalloproteins with X-ray absorption spectroscopy will be a very powerful tool for gaining information on metal centers, especially novel centers with no homologs, when structural information is not immediately forthcoming. While substitution of an incorrect metal such as zinc for iron in heterologous expression in *E. coli* may allow proper folding and give a correct overall structure, the details of the metal center, which may well be critical for understanding the function, will not be correct. Ultimately, this problem will require a return to the native *Pyrococcus* protein, to determine what the ‘correct’ native metal cofactor is. The structural genomics pipeline will, however, provide a wealth of information on novel metalloproteins, and indicate which proteins are the best candidates for further investigation.

The Southeast Collaboratory for Structural Genomics is supported by grants from the National Institutes of Health (GM 62407), the Georgia Research Alliance and the University of Georgia. The authors would also like to thank student

assistants Brian Gerwe, John Mackert and Danny Tran for their help.

## References

- Adams, M. W. W., Dailey, H. A., Delucas, L. J., Luo, M., Prestegard, J. H., Rose, J. P. & Wang, B.-C. (2003). *Acc. Chem. Res.* **36**, 191–198.
- Adams, M. W. W., Holden, J. F., Menon, A. L., Schut, G. J., Grunden, A. M., Hou, C., Hutchins, A. M., Jenney, F. E. Jr, Kim, C., Ma, K., Pan, G., Roy, R., Sapra, R., Story, S. V. & Verhagen, M. F. (2001). *J. Bacteriol.* **183**, 716–724.
- Apiyo, D. & Wittung-Stafshede, P. (2002). *Protein Sci.* **11**, 1129–1135.
- Baneyx, F. (1999). *Curr. Opin. Biotechnol.* **10**, 411–421.
- Cornelis, P. (2000). *Curr. Opin. Biotechnol.* **11**, 450–454.
- Czaja, C., Litwiller, R., Tomlinson, A. J., Naylor, S., Tavares, P., LeGall, J., Moura, J. J. G., Moura, I. & Rusnak, F. (1995). *J. Biol. Chem.* **270**, 20273–20277.
- Daujotyte, D., Vilkaitis, G., Manelyte, L., Skalicky, J., Szyperski, T. & Klimašauskas, S. (2003). *Protein Eng.* **16**, 295–301.
- Degtyarenko, K. (2000). *Bioinformatics.* **16**, 851–864.
- Eidsness, M. K., O'Dell, S. E., Kurtz, D. M. Jr, Robson, R. L. & Scott, R. A. (1992). *Protein Eng.* **5**, 367–371.
- Fiala, G. & Stetter, K. O. (1986). *Arch. Microbiol.* **145**, 56–61.
- Fox, J. D., Routzahn, K. M., Bucher, M. H. & Waugh, D. S. (2003). *FEBS Lett.* **537**, 53–57.
- Giles, N. M., Watts, A. B., Giles, G. I., Fry, F. H., Littlechild, J. A. & Jacob, C. (2003). *Chem. Biol.* **10**, 677–693.
- Hantke, K. (2001). *Curr. Opin. Microbiol.* **4**, 172–177.
- Heinemann, U., Frevert, J., Hofmann, K., Illing, G., Maurer, C., Oschkinat, H. & Saenger, W. (2000). *Prog. Biophys. Mol. Biol.* **73**, 347–362.
- Henricksen, L. A., Umbricht, C. B. & Wold, M. S. (1994). *J. Biol. Chem.* **269**, 11121–11132.
- Holden, J. F., Poole, F. L. II, Tollaksen, S. L., Giometti, C. S., Lim, H., Yates, J.R. III & Adams, M. W. W. (2001). *Comp. Funct. Genom.* **2**, 275–288.
- Holm, R. H., Kennepohl, P. & Solomon, E. I. (1996). *Chem. Rev.* **96**, 2239–2314.
- Jenney, F. E. Jr & Adams, M. W. W. (2001). *Methods Enzymol.* **334**, 45–55.
- Jonasson, P., Liljeqvist, S., Nygren, P.-Å. & Ståhl, S. (2002). *Biotechnol. Appl. Biochem.* **35**, 91–105.
- Kang, D.-K., Jeong, J., Drake, S. K., Wehr, N. B., Rouault, T. A. & Levine, R. L. (2003). *J. Biol. Chem.* **278**, 14857–14864.
- Krishna, S. S., Majumdar, I. & Grishin, N. V. (2003). *Nucl. Acids Res.* **31**, 532–550.
- Li, C., Schwabe, J. W., Banayo, E. & Evans, R. M. (1997). *Proc. Natl. Acad. Sci. USA*, **94**, 2278–2283.
- Liu, C. & Xu, H. (2002). *J. Inorg. Biochem.* **88**, 77–86.
- Makarova, K. S., Aravind, L. & Koonin, E. V. (2002). *Trends Biochem. Sci.* **27**, 384–386.
- Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R. R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S. E., Pagni, M., Peyruc, D., Ponting, C. P., Selengut, J. D., Servant, F., Sigrist, C. J. A., Vaughan, R. & Zdobnov, E. M. (2003). *Nucl. Acids Res.* **31**, 315–318.
- Pedelacq, J. D., Piltch, E., Liang, E. C., Berendzen, J., Kim, C. Y., Rho, B. S., Park, M. S., Terwilliger, T. C. & Waldo, G. S. (2002). *Nat. Biotechnol.* **20**, 927–932.
- Robb, F. T., Maeder, D. L., Brown, J. R., DiRuggiero, J., Stump, M. D., Yeh, R. K., Weiss, R. B. & Dunn, D. M. (2001). *Methods Enzymol.* **330**, 134–157.
- Sambrook, J. & Russell, D. W. (2001). *Molecular Cloning, A Laboratory Manual*, 3rd ed. New York: Cold Spring Harbor Laboratory Press.
- Schmid, B., Ribbe, M. W., Einsle, O., Yoshida, M., Thomas, L. M., Dean, D. R., Rees, D. C. & Burgess, B. K. (2002). *Science*, **296**, 352–356.
- Schut, G. J., Brehm, S. D., Datta, S. & Adams, M. W. W. (2003). *J. Bacteriol.* **185**, 3935–3947.
- Schut, G. J., Zhou, J. & Adams, M. W. W. (2001). *J. Bacteriol.* **183**, 7027–7036.
- Scott, R. A., Shokes, J. E., Cospser, N. J., Jenney, F. E. & Adams, M. W. W. (2005). *J. Synchrotron Rad.* **12**, 19–22.
- Stevens, J. M., Rao Saroja, N., Jaouen, M., Belghazi, M., Schmitter, J.-M., Mansuy, D., Artaud, I. & Sari, M.-A. (2003). *Prot. Exp. Purif.* **29**, 70–76.
- Verhagen, M. F., Menon, A. L., Schut, G. J. & Adams, M. W. W. (2001). *Methods Enzymol.* **330**, 25–30.
- Wittung-Stafshede, P. (2002). *Acc. Chem. Res.* **35**, 201–208.
- Yee, A., Pardee, K., Christendat, D., Savchenko, A., Edwards, A. M. & Arrowsmith, C. H. (2003). *Acc. Chem. Res.* **36**, 183–189.