Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Application of high-throughput technologies to a structural proteomics-type analysis of *Bacillus anthracis*

**K. Au,[a] N. S. Berrow,[a] E. Blagova,[b] I. W. Boucher,[b] M. P. Boyle,[b] J. A. Brannigan,[b] L. G. Carter,[a] T. Dierks,[c] G. Folkers,[c] R. Grenha,[b] K. Harlos,[a] R. Kaptein,[c] A. K. Kalliomaa,[b] V. M. Levdikov,[b] C. Meier,[a] N. Milioti,[b] O. Moroz,[b] A. Müller,[b] R. J. Owens,[a] N. Rzechorzek,[b] S. Sainsbury,[a] D. I. Stuart,[a] T. S. Walter,[a] D. G. Waterman,[b] A. J. Wilkinson,[b] K. S. Wilson,[b] N. Zaccai,[a] Robert M. Esnouf[a]* and Mark J. Fogg[a]***

[a]Division of Structural Biology, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, England, [b]The York Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, York YO10 5YW, England, and [c]Bijvoet Center for Biomolecular Research, NMR Spectroscopy, Utrecht University Padualaan 8, 3584 CH Utrecht, The Netherlands

Correspondence e-mail: robert@strubi.ox.ac.uk, fogg@ysbl.york.ac.uk

Received 10 March 2006
Accepted 21 August 2006

A collaborative project between two Structural Proteomics In Europe (SPINE) partner laboratories, York and Oxford, aimed at high-throughput (HTP) structure determination of proteins from *Bacillus anthracis*, the aetiological agent of anthrax and a biomedically important target, is described. Based upon a target-selection strategy combining 'low-hanging fruit' and more challenging targets, this work has contributed to the body of knowledge of *B. anthracis*, established and developed HTP cloning and expression technologies and tested HTP pipelines. Both centres developed ligation-independent cloning (LIC) and expression systems, employing custom LIC-PCR, Gateway and In-Fusion technologies, used in combination with parallel protein purification and robotic nanolitre crystallization screening. Overall, 42 structures have been solved by X-ray crystallography, plus two by NMR through collaboration between York and the SPINE partner in Utrecht. Three biologically important protein structures, BA4899, BA1655 and BA3998, involved in tRNA modification, sporulation control and carbohydrate metabolism, respectively, are highlighted. Target analysis by biophysical clustering based on pI and hydropathy has provided useful information for future target-selection strategies. The technological developments and lessons learned from this project are discussed. The success rate of protein expression and structure solution is at least in keeping with that achieved in structural genomics programs.

## 1. Introduction

The pan-European project Structural Proteomics In Europe (SPINE; http://www.spineurope.org/) is primarily aimed at fostering collaborations to yield methods developments, particularly those relevant to high-value eukaryotic, especially human, proteins of biomedical importance. However, two SPINE partners, Oxford and York, have pursued a collaborative pilot project on a biomedically important prokaryotic target, *Bacillus anthracis*, as a vehicle for the development of robust technologies for protein expression, crystallization and structure determination. This paper looks at the developments driven by this collaboration and analyses trends that have emerged from the results.

The Oxford Protein Production Facility (OPPF) is funded by the UK Medical Research Council to pilot high-throughput (HTP) protein production and crystallization methods. From the outset, its remit has been to tackle challenging biomedically relevant proteins, but as the technical platform for protein production took shape the need emerged to stress-test the pipeline under true HTP conditions. At the same time, the

York laboratory wanted to use its involvement in SPINE to investigate more automated methodologies and already had a long-standing interest and well developed expertise in the cloning, expression and crystallization of proteins from the non-pathogenic soil-dwelling bacterium *B. subtilis* (Lewis *et al.*, 1998, 2002; Cladière *et al.*, 2004), with the main areas of interest including sporulation and its control, ABC transport systems and two-component signal transduction. To meet the goals of both laboratories, it was decided to pursue a collaborative study of proteins from *B. anthracis*, a close genetic relative of *B. subtilis*, since it is of clear biomedical importance and the complete genome sequences of both *B. subtilis* [4100 predicted open reading frames (ORFs); Kunst *et al.*, 1997] and *B. anthracis* (5311 predicted ORFs; Read *et al.*, 2003) were available. *B. anthracis*, a highly virulent Gram-positive endospore-forming bacterium, is the aetiological agent of anthrax (Mock & Fouet, 2001), a member of the '*B. cereus* group' of group 1 bacilli (Radnedge *et al.*, 2003) and can be isolated from soil environments as dormant spores able to survive for decades (Manchee *et al.*, 1981; Turnbull, 2002). The longevity of *B. anthracis* spores and relative ease of production gives them potential as a bioweapon (Inglesby *et al*., 2002). Anthrax is normally a disease of grazing herbivores, but it can also infect other mammals including humans. Infections are acquired through the inhalation (inhalation anthrax) or ingestion (gastrointestinal anthrax) of *B. anthracis* spores, by spore entry through breaks in the skin (cutaneous anthrax) and even from insect bites (Mock & Fouet, 2001). The genome sequences of *B. anthracis* (Read *et al.*, 2003) and *B. cereus* (Ivanova *et al.*, 2003) show clear differences from that of *B. subtilis* (Anderson *et al.*, 2005), consistent with the different ecological niches inhabited by each organism. These differences combined with comparative functional and structural analyses of proteins common across the *B. cereus* group and those unique to pathogenic species could lead to new non-antibiotic therapies.

## 2. Materials and methods

### 2.1. Target-selection strategies

Development of bioinformatics tools for target selection formed an early part of the HTP developments at Oxford and the OPAL and OPTIC resources are described in Albeck *et al.* (2006). The OPTIC database was populated with the 5311 predicted *B. anthracis* ORFs (Read *et al.*, 2003), a resource used by both Oxford and York to inform the target-selection process. York enhanced their selection with the 'YSBL Structural Genomics Resource' (Rodrigues & Hubbard, 2003), populated with all *B. anthracis* ORFs and exploiting many of the same resources as OPAL and OPTIC.

For York, the target-selection strategy combined developing ongoing research interests with deliberate selection of uncomplicated targets to assess the potential of HTP methodologies. An initial set of 48 targets for use in HTP structural proteomics was selected as low-hanging fruit by the following criteria: (i) molecular weight <50 kDa, (ii) candidate for molecular replacement, (iii) not part of a complex, (iv) contains no signal peptides or transmembrane regions, (v) predicted to be soluble and ordered according to *PONDR* (Romero *et al.*, 2001). Targets that were members of families previously studied by York were particularly favoured. A second set of 96 targets was selected on the basis of three specific criteria: (i) 24 targets based on the identification of genes present in pathogenic *Bacillus* species (*B. anthracis* and *B. cereus*) but not in the genetically close non-pathogenic organism *B. subtilis*; (ii) 27 targets having significant biological interest, having no molecular-replacement model and/or being essential to the viability of the organism [primarily conserved hypothetical protein homologues of *B. subtilis* proteins identified from the Database of Essential Genes (DEG; Zhang *et al.*, 2004)]; (iii) 45 targets selected on bioinformatics criteria similar to those above but including disorder evaluation with *RONN* (Yang *et al.*, 2005; see also Esnouf *et al.*, 2006). Furthermore, sets of 10–20 targets matching existing York research interests were selected for individual study, such as hypothetical histidine kinases and representatives of the glycoside hydrolase, carbohydrate esterase and glycosyltransferase families that had not as yet been fully characterized.

Since the primary goal for the Oxford *B. anthracis* study was to test the OPPF pipeline, a set of 48 proteins from well conserved protein families were selected since these have been hypothesized to be most amenable to structural study, presumably because they represent the physically stable core machinery of the cell (O. Herzberg, personal communication). The initial procedure involved selecting a set of 23 diverse pathogenic bacteria (including both Gram-positive and Gram-negative bacteria for which full genome sequences were available[1]) followed by an exhaustive all-against-all *BLAST* alignment (Altschul *et al.*, 1990) for every protein sequence predicted from these genomes. *TribeMCL* (Enright *et al.*, 2002) was then used to cluster the significant hits (*BLAST E* value <0.1) into families and *B. anthracis* proteins were considered further if they belonged to families conserved across 22 (239 families) or 23 (203 families) of the bacterial genomes. The list of possible targets was further filtered in a similar process to that employed by the York group based on size, prediction of transmembrane regions and signal peptides and possibility of structure solution by molecular replacement. A target-selection meeting finalized the choice of targets and included a few 'challenging' targets such as those for which no molecular-replacement solution was available and four conserved hypothetical proteins. As the Oxford pipeline developed, these targets were revisited and some construct redesign was attempted for previously unsuccessful targets. Finally, a second set, this time of 78 targets, was selected based

[1] *Bacillus anthracis*, *Bacillus subtilis*, *Brucella melitensis*, *Campylobacter jejuni*, *Chlamydia muridarum*, *Chlamydophila pneumoniae*, *Clostridium tetani*, *Deinococcus radiodurans*, *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter hepaticus*, *Lactobacillus plantarum*, *Listeria innocua*, *Mycobacterium tuberculosis*, *Mycoplasma pneumoniae*, *Neisseria meningitides*, *Pseudomonas aeruginosa*, *Salmonella typhi*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Streptomyces coelicolor*, *Vibrio cholerae*, *Yersinia pestis*.

on biological interest while relaxing the family conservation criterion slightly. For some targets several constructs were attempted in parallel, yielding a total of 96 constructs.

### 2.2. Experimental strategy at York

York has devised a semi-automated ligation-independent cloning (LIC-PCR or LIC; Aslanidis & de Jong, 1990) strategy with *Escherichia coli* expression and non-cleavable N-terminal $His_6$ tags suitable for the high-throughput cloning and over-expression of bacterial targets (M. Fogg, manuscript in preparation; Alzari *et al.*, 2006). The cloning and expression vector, pET-YSBLIC, was engineered for LIC by directed mutagenesis of pET-28a (Novagen) that added an N-terminal tag (MGSSHHHHHH-) directly onto any expressed protein cloned into the vector. A small number of targets were cloned into unmodified pET-28a by conventional restriction/ligation methods.

The typical protein-production strategy is described in Alzari *et al.* (2006). Target coding sequences were amplified from *B. anthracis* genomic DNA (Ames strain) by PCR, cloned into pET-YSBLIC and transformed into *E. coli* NovaBlue cells (Novagen). Target clones identified by colony PCR were transformed into *E. coli* BL21 (DE3), B834 (DE3) and/or Rosetta2 (DE3) for protein-overexpression trials and scale-up. Following purification, protein concentrations were adjusted to 15–20 mg $ml^{-1}$ for crystallization trials.

Crystallization screening used a Mosquito nanolitre liquid-handling robot (TTP Labtech), standard sparse-matrix conditions and CrystalQuick 96-well crystallization plates (Greiner Bio-One). Where possible, crystals from these trials were tested for X-ray diffraction in-house. However, successful crystallization trials were usually scaled up and/or used as the starting point for optimization/additive screening, either by hand or using the Mosquito robot, to improve crystal size and quality. After in-house characterization, the best crystals were shipped pre-frozen to synchrotrons (ESRF, Grenoble or SRS, Daresbury) for data collection. Most structures were solved by molecular replacement.

Nuclear magnetic resonance (NMR) experiments have augmented crystallographic studies (see also AB *et al.*, 2006). Initially, $^{15}N$-labelled *B. anthracis* proteins were screened by recording their $^{15}N$-HSQC fingerprint NMR spectra as described previously (Folkers *et al.*, 2004). However, two targets that failed to yield protein crystals, BA1655 and BA5174, proved suitable for structure determination by solution NMR. NMR conditions were optimized for both proteins to increase concentration and stability while reducing both the pH and overall salt conditions. Complete sets of NMR spectra for assignment and structure elucidation were recorded on Bruker Avance spectrometers operating at 700 and 900 MHz $^1H$ frequency. These high field strengths offered crucial gains in resolution for both proteins (see AB *et al.*, 2006), which display the rather low dispersion of $H^N$ and $H^\alpha$ protons typical of all-helical proteins (see York example structure below).

### 2.3. Experimental strategy at Oxford

The OPPF protein-production pipeline also uses ligation-independent methods. The strategy has evolved during the course of this *B. anthracis* study partly as a result of the lessons learned from it and these methods are outlined in Alzari *et al.* (2006). Initially, experiments focused on Gateway technology using the pET15G expression vector (Luan *et al.*, 2004), which adds a 3C-cleavable N-terminal $His_6$ tag but also a translation of the *attB* site giving a total tag length of 38 amino acids that appeared to restrict protein expressibility/solubility. Subsequent work used the pDEST14 expression vector (Invitrogen), which only adds a minimal non-cleavable N-terminal $His_6$ tag (MAHHHHHH-) and gave much better levels of soluble expression. Work on the set of 78 targets has been based on the ligation-independent In-Fusion system (Clontech) using in-house modifications of the pET-28a and pTRIEX2 (Novagen) expression vectors (namely pOPINB and pOPINF, respectively) that can add either N-terminal or C-terminal $His_6$ purification tags (see Alzari *et al.*, 2006; Berrow *et al.*, manuscript in preparation). Target coding sequences were amplified by PCR directly from *B. anthracis* genomic DNA (supplied by York) or indirectly using the GenomiPhi procedure (Dean *et al.*, 2001). Cloning and small-scale expression trials followed the methods described in Alzari *et al.* (2006), with expression scale-up in either *E. coli* B834 (DE3) or Rosetta (DE3) cells. Purified proteins were concentrated for crystallization trials, with pre-crystallization tests (Hampton) being used to suggest appropriate protein concentrations.

Crystallization screening followed the standard Oxford sitting-drop protocols for screening and optimization (Walter *et al.*, 2003, 2005; Brown *et al.*, 2003) using sparse-matrix conditions and CrystalQuick 96-well crystallization plates (Greiner Bio-One). Crystals taken from these nanolitre screens and optimizations were used directly for data collection, with crystals shipped to synchrotrons (ESRF, Grenoble or SRS, Daresbury) in plates and mounted at the beamline. Most structures were solved by molecular replacement, although selenium multiwavelength anomalous dispersion (MAD) analysis at BM14, Grenoble was used in some cases.

## 3. Results

### 3.1. Structures determined by York

Table 1 summarizes the progress made by York. The parallel experiments on sets of 48 and 96 targets are at different stages, with work on the 48-target set largely completed, whereas work on the set of 96 is ongoing. The high-resolution limits of collected data sets ranged from 1.6 to 3.5 Å. A total of 21 structures have been determined so far: ten from the set of 48 targets, two from the set of 96 targets, five from other targets, two structures with ligands and two determined by NMR.

The efficacy of the LIC method is clearly demonstrated as 46/48 target ORFs were cloned into the pET-YSBLIC vector after just two rounds of cloning, with 33/46 (72%) expressed in soluble form and in sufficient yield to enter crystallization

**Table 1**
Progress on different *B. anthracis* target sets by the York and Oxford laboratories.

Work on the first sets of 48 targets for each laboratory is largely finished and gives an indication of potential eventual success rates. Work on the other target sets is ongoing.

| Progression | York set of 48 | York set of 96 | York others | Oxford set of 48 | Oxford set of 78 | Total progress |
|---|---|---|---|---|---|---|
| Selected | 48 | 96 | 89 | 48 | 78 | 359 |
| PCR | 48 | 94 | 88 | 47 | 76 | 353 |
| Expressed | 43 | 54 | 67 | 43 | 28 | 235 |
| In crystallization | 33 | 38 | 40 | 41 | 9 | 161 |
| Crystallized | 32 | 23 | 11 | 22 | 6 | 94 |
| Data sets | 12 | 6 | 9 | 17 | 5 | 49 |
| Structures solved | 10 | 2 | 5 | 15 | 4 | 36 |
| Structures with ligands | 2 | — | — | 2 | 2 | 6 |
| NMR structures | — | — | 2 | — | — | 2 |
| Total structures | 12 | 2 | 7 | 17 | 6 | 44 |

trials following protein purification. Diffraction-quality crystals were obtained for 18 of these proteins, yielding 12 data sets with ten new *B. anthracis* protein structures solved, two of which were also solved with bound ligands. The ongoing work with the more challenging set of 96 targets has yielded single-pass statistics with a subsequent reduction in the number (72/96; 75%) of cloned ORFs, as only a single round of cloning, expression and purification has been carried out. However, soluble expression statistics compare favourably with the first set, as 38/54 (70%) of targets so far tested have produced soluble protein expression suitable for crystallization trials. To date, this has resulted in six diffraction data sets and two new structures. All other targets, progressed in smaller groups by individual graduate students, are at various stages of completion. This disparate target group comprises subsets selected on the basis of specific biological interest with little weight given to overall tractability; consequently, it comprises the most challenging set.

The discrepancy between data sets collected and solved structures (Table 1) is a consequence of problems of crystal or data quality, including problems such as twinning, limited resolution and data incompleteness. The majority of constructs used and hence proteins crystallized had the non-cleavable $His_6$ purification tag (MGSSHHHHHH-) attached; consequently, in order to assess whether poor-quality crystals were a result of the presence of this tag, a vector pET-YSBLIC3C was developed containing a 3C protease cleavage site for tag removal. An assessment of whether this method significantly improves crystal quality is ongoing, although based on the Oxford experiences described below we would expect some benefit.

### 3.2. Examples

The tRNA-modifying enzyme ThiI from *B. anthracis* (BA4899) was selected as a target for structural determination as it contains a THUMP domain, a predicted RNA-binding module found in various RNA-modifying enzymes throughout the three domains of life (Aravind & Koonin, 2001). ThiI is responsible for the formation of 4-thiouridine at position 8 of tRNA, a common modification in prokaryotes. This base forms a covalent photo-cross-link with a cytosine at position 13 of tRNA when irradiated with near-UV light, thus providing a sensor for such exposure and a trigger for the subsequent cellular response (Ramabhadran & Jagger, 1976). Following cloning (pET-YSBLIC), expression and crystallization, the structure of ThiI was solved (Waterman *et al.*, 2006) in complex with AMP to 2.5 Å resolution by MAD analysis of crystals of a selenomethionine-substituted form of the protein (Fig. 1*a*) using data collected on beamline BM14 at the ESRF, Grenoble. The structure reveals the THUMP fold to be unrelated to that of previously characterized RNA-binding domains. In ThiI, the THUMP domain is accompanied by an N-terminal ferredoxin-like domain. Analysis of conserved exposed residues indicates that the tRNA-binding surface is likely to be formed from both domains; however, a model of the interaction remains elusive. The modified uridine at position 8 is buried within the core of tRNA in its canonical L-shaped conformation, implying that structural changes to the tRNA molecule are necessary to expose it to the PP-loop pyrophosphatase active site of the enzyme.

Two targets, BA1655 and BA5174, were identified as members of the Spo0E-like phosphatase family by amino-acid sequence comparison with the *B. subtilis* proteins Spo0E, YnzD and YisI. In *B. subtilis*, Spo0A is the key regulator of the sporulation phosphorelay. The threshold concentration of Spo0A~P required for sporulation is achieved by the activity of sporulation sensor kinases. Spo0E-like phosphatases counter this by dephosphorylating Spo0A~P, thereby inhibiting sporulation (Ohlsen *et al.*, 1994; Perego *et al.*, 1994). Whilst crystallization trials with BA1655 and BA5174 failed, both proteins were considered suitable candidates for structure determination by NMR, being small (7.5 and 7.9 kDa, respectively) and highly soluble. The structures were solved in collaboration with the SPINE partner in Utrecht. Initial $^{15}$N-HSQC spectra demonstrated a good dispersion typical of $\alpha$-helical proteins, with BA1655 exhibiting greater stability than BA5174, the latter requiring additional temperature and concentration optimization. Complete sets of NMR spectra led to structure elucidation of BA1655 (Grenha *et al.*, 2006, in the press) comprising a dimer in a four-helix bundle and consisting of two pairs of helices connected by a tight turn packed in a head-to-tail manner (Fig. 1*b*), whereas BA5174 is monomeric and comprises a similar pair of antiparallel $\alpha$-helices. These structures indicated that crystallization difficulties probably arose from disorder at the C-terminus. These tails were too short (13 and 16 residues for BA5174 and BA1655, respectively) to be reliably predicted as disordered by the algorithms used during target selection.

### 3.3. Structures determined by Oxford

A summary of Oxford progress on the first 48 targets and the current (early) state of progress with the second 78 targets is provided in Table 1. Overall, 23 structures have been determined (including complexes with small molecules) for 19

different target proteins with high-resolution limits ranging from 1.6 to 3.3 Å, but the success rates for the different protocols explored has varied greatly.

For the pET15G expression vector (long, cleavable N-terminal tag) cloning from genomic DNA was successful (41/48; 85%), but the fraction of targets yielding soluble expression in screening was low (13/41; 32%) and even for these the expression levels were often less than 5 mg $l^{-1}$. Expression of only six targets could be scaled up successfully for crystallization trials and these trials were set up using both cleaved and uncleaved proteins. Diffraction-quality crystals were obtained with three of the cleaved proteins and all yielded structures. In addition, one target was successfully cocrystallized with substrate and product, and the structure determined. The uncleaved protein yielded crystals for only one protein and these were of poor quality. Work on the set of 48 targets was repeated using the pDEST14 expression vector to give a combined coverage of targets of 47/48 (98%). The pDEST14 vectors (short, uncleavable N-terminal purification tag) were screened for soluble expression on a small scale (1 ml) using both IPTG and autoinduction (Studier, 2005) protocols with generally similar results for well expressing proteins, although for poorly expressing proteins autoinduction often outperforms the IPTG protocol. Soluble expression levels were generally much higher from the pDEST14 vectors compared with the equivalent pET15G vectors (typical yields in excess of 5 mg $l^{-1}$) and a total of 33/44 (75%) targets entered into crystallization trials. Crystals were grown for 18/33 (55%) of these targets; X-ray data sets were collected for 14 and yielded 11 structures, including one overlap (at lower resolution) with a target from the pET15G round.

One of the failed targets was used as part of an initial test of the In-Fusion technology (Clontech) and this yielded good soluble expression (>5 mg $l^{-1}$) and a structure at 2.4 Å resolution, prompting further evaluation and eventual adoption of this technology. In-Fusion-based work with some of the remaining recalcitrant targets from the initial 48 targets yielded crystals of three new constructs, two medium-resolution (3–3.5 Å) data sets and one structure. As a result of these successes, In-Fusion technology is also being used for the new set of 78 targets (some of which are being processed with multiple constructs, giving a total of 96 constructs), of which 28 have to date yielded soluble expression.

### 3.4. Example

The structure determination of *B. anthracis* D-ribulose-5-phosphate epimerase (BA3998) showed several interesting features (Carter *et al.*, in preparation). The protein expressed in soluble form with the long 3C protease-cleavable tag and was the only target for which (poor-quality) crystals could be obtained with the tagged protein. After cleavage, however, good crystals were obtained and X-ray data were collected extending to 2.2 Å resolution for a crystal belonging to space group $P4_12_12$ on Station 14.1 at the SRS, Daresbury. An attempted soak with D-ribulose-5-phosphate was ultimately unsuccessful, but improved native data extending to 2.05 Å were collected on beamline ID14EH1 at the ESRF, Grenoble. The structure of the substrate-free epimerase was solved by molecular replacement to reveal a hexamer arranged as a triangular prism (Fig. 1*c*; $R$ factor = 0.208; $R_{\rm free}$ = 0.236). Slowly growing crystals were obtained from cocrystallizations with the same substrate and after an initial 2.9 Å resolution data set confirmed the presence of bound substrate and/or product, a data set extending to 2.3 Å resolution was recorded on beamline ID14EH2 at the ESRF from a cocrystal that took more than six months to appear and belonged to space group $P2_12_12_1$. This structure was solved by molecular replacement starting from the substrate-free structure and revealed D-ribulose-5-phosphate (probably mixed with the product, D-xylulose-5-phosphate) bound to differing extents in the six active sites of the hexamer along with catalytic zinc ions ($R$ factor = 0.211; $R_{\rm free}$ = 0.266). These data are the first for a substrate/product-bound epimerase and allow us to modify and update the previously suggested model for epimerase catalysis (Kopp *et al.*, 1999; Jelakovic *et al.*, 2003).

## 4. Discussion

### 4.1. Cloning strategies

Traditional methods of cloning using restriction enzymes require tailoring of the protocol to each individual target and hence are not amenable to HTP approaches. This has stimulated interest in ligation-independent methods where only the primers vary between targets. York has developed a custom LIC-based approach, whereas Oxford has concentrated on commercial technologies, initially Gateway and more recently In-Fusion. All methods have proven amenable to parallelization and success rates are high (Table 1) and relatively independent of target size. Oxford has adapted the cloning protocols to work effectively on two robotic liquid-handling stations, the BioRobot 8000 (Qiagen) and the MWG Theonyx (Aviso GmbH), whereas York are in the process of adapting their protocols for robotic liquid handling on a Freedom EVO (Tecan) as part of their newly established High-Throughput Expression Laboratory (HiTEL). Primer design is also relatively automated; for example, the OPPF primer-design interface, *OPINE*, which is built around the Primer3 package, provides a direct link between the target-selection software and the LIMS (Albeck *et al.*, 2006). *OPINE*, along with the Web Primer resource (http://seq.yeastgenome.org/cgi-bin/web-primer), was used remotely by York for early primer-design requirements, although a new primer-design program developed in-house is now used for all LIC and LIC3C primers. One advantage of the York and In-Fusion protocols is that the primers are shorter than those required to produce the same short-tagged product with Gateway technologies; furthermore, only one experimental reaction is required, resulting in significant cost and time savings.

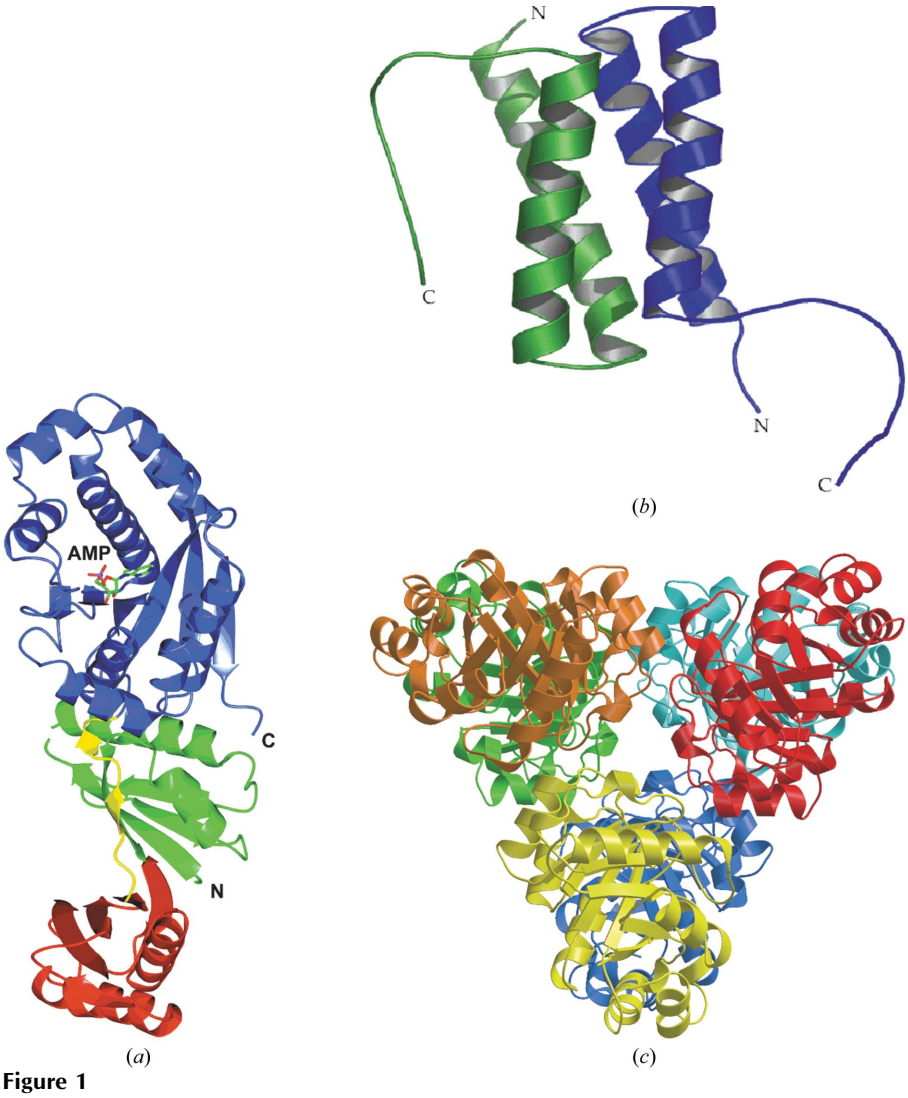
### 4.2. Expression and purification strategies

Not surprisingly, for bacterial proteins high levels of soluble expression are routinely observed by York and Oxford using a

variety of *E. coli* strains. Purification strategies in both laboratories are now based on ÄKTA XPress and Explorer 3D systems (GE Healthcare) and are found to be very effective. Where the Oxford laboratory requires cleavage of N-terminal purification tags, this is routinely performed with a His-tagged 3C protease, which is now seamlessly integrated into the HTP context and works with high efficiency. The effect of different tags in expression is discussed in §4.5.

### 4.3. Crystallization and crystal quality

Oxford has helped pioneer the use of small-volume (100 + 100 nl drop) crystallization trials (Walter *et al.*, 2003, 2005; Brown *et al.*, 2003) and now has two Cartesian Technologies MicroSys MIC400 (Genomic Solutions Ltd) robots in heavy use. York has deployed the Mosquito (TTP Labtech) robot. Both laboratories have observed an increase in success rate on reduction of drop volume. Oxford rapidly moved towards a strategy of optimization using small-volume drops and has developed standard procedures which have now been implemented on liquid-handling robots and are very effective, avoiding the need to translate conditions to a larger drop format (Walter *et al.*, 2005). Crystals from small drops are used directly for data collection, although these crystals are frequently too small to be screened using in-house X-ray facilities. Thus, Oxford has tended to take crystallization trays to synchrotrons and explore cryoprotection strategies on the beamline, an effective but labour-intensive approach. In contrast, York has tended to scale up crystallization-drop volumes during the optimization process and as a result the crystals have usually been large enough for in-house screening; nevertheless, four of the 21 York structures were solved using crystals recovered directly from 96-well screens. The York strategy allows scaled-up in-house crystals to be shipped frozen. The effect of the purification tag on crystal quality is discussed in §4.5.




(*a*) (*b*) (*c*)

**Figure 1**
Selected structures of *B. anthracis* proteins solved by the York and Oxford laboratories. (*a*) A monomer of ThiI (BA4899) shown as a ribbon representation. The N-terminal ferredoxin-like domain is coloured green and forms a continuous $\beta$-sheet surface with the THUMP domain (red). The enzymatic PP-loop pyrophosphatase domain is coloured blue and shows the position (as a stick representation) of AMP bound at the active site. The glycine-rich linker joining the THUMP domain to the PP-loop domain is coloured yellow. (*b*) A ribbon representation of the lowest energy structure (PDB code 2bzb) of the Spo0E-like phosphatase target BA1655 showing the individual monomers in green and blue. (*c*) A ribbon diagram showing the hexameric structure of unliganded D-ribulose-5-phosphate epimerase (BA3998). Individual chains are coloured red, orange, yellow, green, light blue and dark blue. The hexameric structure of the protein with bound substrate has also been determined.

### 4.4. Structure-determination strategies

Most of the targets studied by both York and Oxford were amenable to structure determination by molecular replacement. In general, structure solutions proceeded in a routine manner, with automated molecular-replacement packages, such as *CaspR* (Claude *et al.*, 2004) and the CCP4/eHTPX-supported development of *MrBUMP* (Keegan & Winn, manuscript in preparation; Bahar *et al.*, 2006) often proving successful. Indeed, several of the *B. anthracis* data sets were used to help in the development and testing of *MrBUMP*. Data for one of the Oxford *B. anthracis* targets proved particularly challenging for automated methods and have helped to drive further developments. These data (Bahar *et al.*, 2006) are somewhat unusual in that they are from a highly mosaic crystal (>2°) that nevertheless diffracts to high resolution (1.5 Å).

### 4.5. The effect of the choice of tag on overall success

Both York and Oxford routinely use vectors that introduce short $His_6$ purification tags. York has concentrated almost exclusively on the non-cleavable N-terminal MGSSH-HHHHH- tag, with a 3C protease-cleavable tag version (pET-YSBLIC3C) currently undergoing assessment. Oxford continues to explore cleavable and non-cleavable N-terminal tags as well as cleavable C-terminal tags since this has been found to increase the success rates for 'difficult' (*i.e.* non-bacterial) targets as part of a multi-construct approach (Siebold *et al.*, 2005). Experience at both laboratories is that longer tags adversely affect the ability to crystallize. From six Oxford targets that produced satisfactory levels of soluble expression with a long tag, only one gave crystals which were of poor quality while the tag was still attached. After cleavage, good-quality crystals could be grown for three of these targets. In general, it appears that the short N-terminal non-cleavable tags used extensively by both York and Oxford have not markedly impaired success rates for crystallization; however, a quantitative assessment of the effect on crystal quality of the short tag introduced by the pET-YSBLIC vector is ongoing in York. Oxford has already generated a cohort of results ($N = 13$) to help address the question of whether removing the His tag is necessary. In four cases new or improved crystals were obtained after tag cleavage, whilst in only two cases were crystals better with an uncleaved tag. On the basis of these results, the Oxford strategy is now to work with cleavable tags that are as short as possible (both N-terminal 3C cleavable tags and C-terminal carboxypeptidase cleavable tags). It is our experience that the presence of the cleavage site does not of itself significantly reduce protein expression or solubility compared with a minimal non-cleavable tag.

### 4.6. Are target-selection strategies justified?

Oxford achieved a success rate of 31% for 48 targets by repeating efforts and refining protocols, whereas York achieved a success rate of 21% from the completed set of 48 targets but made fewer attempts at target rescue. Groups of proteins within the *B. anthracis* targets showed particular characteristics. For example, of the 111 York targets entered into crystallization trials, 19 (17%) represented TIGR-annotated (http://www.tigr.org) functional categories involved in purine, pyrimidine, nucleoside and nucleotide metabolism, the largest single category of proteins in crystallization. All 19 of these targets produced crystals (of varying quality) with nine leading to structure solution, making this family particularly crystal-friendly. In contrast, the tRNA synthetases tackled by Oxford proved extremely difficult to crystallize

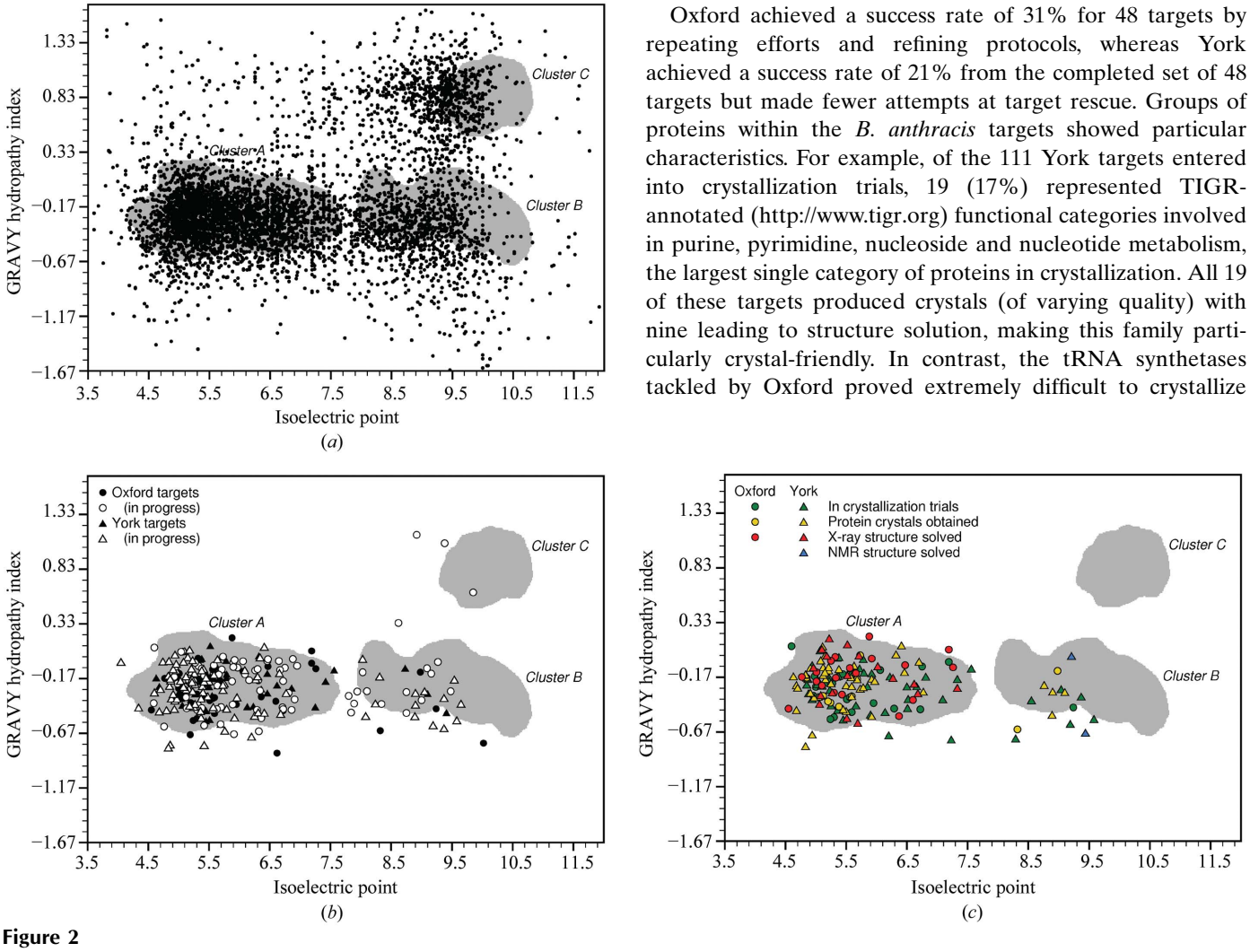



**Figure 2**
Scatter plots of grand average of hydropathy (GRAVY; Kyte & Doolittle, 1982) against pI for proteins from *B. anthracis*. The background shading of each plot shows the clusters (A, B and C) identified from an analysis of *T. maritima* (Canaves *et al.*, 2004). (*a*) A plot of proteins for all predicted *B. anthracis* ORFs. (*b*) A plot of the proteins selected by the York (triangles) and Oxford (circles) laboratories for HTP studies. The largely complete sets of 48 targets are shown as solid shapes, while those for which work is still in progress are shown as outline shapes. (*c*) A plot of targets for which significant progress has been made at York (triangles) and Oxford (circles). Red (X-ray) and blue (NMR) shapes show structures that have been determined; yellow shapes show proteins which have not had structures determined but have yielded crystals; green shapes show proteins which have not given crystals but been expressed solubly in sufficient quantities to enter crystallization trials.

despite giving adequate yields of soluble protein and only one structure (at 2.8 Å resolution) was obtained. This experience reflects the general difficulty of working with tRNA synthetases. As exemplified by the tRNA synthetases, where several members of a single protein family were tackled in Oxford, at least one structure was obtained for every family for which multiple members were targeted (six families)[2] supporting the strategy of selecting paralogues in order to increase overall success. However, it is interesting to note that the several conserved hypothetical proteins that were included because they are widely conserved across bacteria proved difficult. In fact, only one out of four has yielded a structure to date. Finally, the *B. anthracis* targets demonstrate that the use of disorder-prediction methods, such as *RONN* and *PONDR* (see Esnouf *et al.*, 2006) have utility; an example of this is provided by work on a conserved hypothetical protein (BA0541) where redesign of the construct, omitting 18 residues at the N-terminus, yielded a 3.3 Å structure. This structure was the only hypothetical structure solved, but it is notable that unlike the other hypothetical proteins, it belonged to an established fold.

### 4.7. Comparison with a structural genomics study of *Thermotoga maritima*

The combined York/Oxford study of proteins from *B. anthracis* has been on a much smaller scale than the full-genome study of *T. maritima* performed by the JCSG (http://www.jcsg.org). For that study, 1878 ORFs were cloned and have so far resulted in 131 structure determinations. An analysis of the correlates of success with physico-chemical properties has been made for the *T. maritima* project (Canaves *et al.*, 2004) and it appears that some of the findings apply to our *B. anthracis* cohort. In particular, when analysed in terms of the calculated pI and grand average of hydropathy index, the open reading frames of *B. anthracis* group in a similar way to those of *T. maritima* (Fig. 2*a*). However, the C cluster is sparsely populated and the B cluster does not extend to such high pI values. Although these clusters were not considered formally in target selection, it can be seen that both York and Oxford tended strongly to select targets within clusters A and B (Fig. 2*b*). In line with the findings of the JCSG, the targets that led to successful structure determinations were far more likely to come from cluster A (Fig. 2*c*). It is intriguing to note that York did manage to solve two structures from cluster B by NMR analysis. Such plots may therefore be useful as a direct aid to target selection and construct design. Overall, the percentage of targets selected by York or Oxford that yielded structures was higher than that obtained by the JCSG for their comprehensive attack on *T. maritima*, suggesting that the Oxford and York target selection strategies used were of value in eliminating intractable targets.

[2] Families giving a structure: tRNA synthetases, methionine aminopeptidases [BA0132 (structure), BA1590, BA5601], alanine racemase [BA0252 (structure), BA2079], pyruvate kinase [BA3382 (structure), BA4843 (structure)], glyceraldehyde-3-phosphate dehydrogenase [BA4827 (structure), BA5369 (structure)], enoyl CoA hydratase/isomerase [BA0894, BA2356, BA2551, BA3583 (structure)].

## 5. Conclusions

The York and Oxford studies on *B. anthracis* have yielded much more than just structures, although the ∼30% success rate they have achieved alone makes them very successful studies. They have provided a test bed for many of the technological developments in these two laboratories, developments which have been shared across Europe, with dissemination largely enabled by the SPINE project. The work has driven the development of much improved vectors for ligation-independent cloning strategies, has helped to optimize expression and purification protocols, has further validated the HTP nanolitre-scale crystallization methods, has demonstrated the utility of the target-selection tools developed at the start of SPINE and has illustrated ways in which these can be further improved, *e.g.* by the more systematic use of disorder-prediction methods and by considering the clusters in hydropathy and pI. The requirement for conservation across a series of genomes imposed by the OPPF selection criteria also appears to have been beneficial in eliminating difficult proteins, but excludes proteins characteristic of *B. anthracis* that may play an important role in pathogenesis.

Structural genomics studies have provided an enormous impetus for methods development in structural biology and few laboratories are now untouched by their effects. The emphasis in SPINE has been to apply these methods to systems of biological interest, the ultimate aim being to solve significant problems more effectively by the use of HTP and parallel technologies. However, the experience with *B. anthracis* makes it clear that novel technologies should be benchmarked on more tractable targets to reveal technological inadequacies that might otherwise remain hidden within the high attrition rate observed with more challenging problems.

We would like to thank Professor Colin Harwood (Newcastle) for providing genomic DNA. We also thank the staff of BM14 (Grenoble), the SRS (Daresbury), the ESRF and EMBL (Grenoble) for access to synchrotrons and help with data collection. This work was supported by the European Commission as part of SPINE (Structural Proteomics In Europe) contract No. QLG2-CT-2002-00988 under the Integrated Programme 'Quality of Life and Management of Living Resources'. The OPPF is also funded by the Medical Research Council UK.

## References

AB, E. *et al.* (2006). *Acta Cryst.* D**62**, 1150–1161.
Albeck, S. *et al.* (2006). *Acta Cryst.* D**62**, 1184–1195.
Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.
Alzari, P. M. *et al.* (2006). *Acta Cryst.* D**62**, 1103–1113.
Anderson, I. *et al.* (2005). *FEMS Microbiol. Lett.* **250**, 175–184.
Aravind, L. & Koonin, E. V. (2001). *Trends Biochem. Sci.* **26**, 215–217.

Aslanidis, C. & de Jong, P. J. (1990). *Nucleic Acids Res.* **18**, 6069–6074.
Bahar, M. *et al.* (2006). *Acta Cryst.* D**62**, 1170–1183.
Brown, J. *et al.* (2003). *J. Appl. Cryst.* **36**, 315–318.
Canaves, J. M., Page, R., Wilson, I. A. & Stevens, R. C. (2004). *J. Mol. Biol.* **344**, 977–991.
Cladière, L., Blagova, E., Levdikov, V. M., Brannigan, J. A., Seror, S. J. & Wilkinson, A. J. (2004). *Acta Cryst.* D**60**, 329–330.
Claude, J. B., Suhre, K., Notredame, C., Claverie, J. M. & Abergel, C. (2004). *Nucleic Acids Res.* **32**, W606–W609.
Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. (2001). *Genome Res.* **11**, 1095–1099.
Enright, A. J., van Dongen, S. & Ouzounis, C. A. (2002). *Nucleic Acids Res.* **30**, 1575–1584.
Esnouf, R. M., Hamer, R., Sussman, J. L., Silman, I., Trudgian, D., Yang, Z.-R. & Prilusky, J. (2006). *Acta Cryst.* D**62**, 1260–1266.
Folkers, G. E., van Buuren, B. N. M. & Kaptein, R. (2004). *J. Struct. Funct. Genomics*, **5**, 119–131.
Grenha, R., Rzechorzek, N., de Jong, R. N., AB, E., Diercks, T., Truffault, V., Ladds, J. C., Fogg, M. J., Bongiorni, C., Perego, M., Kaptein, R., Wilson, K. S., Folkers, G. E. & Wilkinson, A. J. (2006). In the press.
Inglesby, T. V., O'Toole, T., Henderson, D. A., Bartlett, J. G., Ascher, M. S., Eitzen, E., Friedlander, A. M., Gerberding, J., Hauer, J., Hughes, J., McDade, J., Osterholm, M. T., Parker, G., Perl, T. M., Russell, P. K. & Tonat, K. (2002). *J. Am. Med. Assoc.* **287**, 2236–2252.
Ivanova, N. *et al.* (2003). *Nature (London)*, **423**, 87–91.
Jelakovic, S., Kopriva, S., Suss, K. H. & Schulz, G. E. (2003). *J. Mol. Biol.* **326**, 127–135.
Kopp, J., Kopriva, S., Suss, K. H. & Schulz, G. E. (1999). *J. Mol. Biol.* **287**, 761–771.
Kunst, F. *et al.* (1997). *Nature (London)*, **390**, 249–256.
Kyte, J. & Doolittle, R. F. (1982). *J. Mol. Biol.* **157**, 105–132.
Lewis, R. J., Brannigan, J. A., Offen, W. A., Smith, I. & Wilkinson, A. J. (1998). *J. Mol. Biol.* **283**, 907–912.
Lewis, R. J., Scott, D. J., Brannigan, J. A., Ladds, J. C., Cervin, M. A., Spiegelman, G. B., Hoggett, J. G., Barak, I. & Wilkinson, A. J. (2002). *J. Mol. Biol.* **316**, 235–245.
Luan, C. H., Qiu, S., Finley, J. B., Carson, M., Gray, R., Huang, W., Johnson, D., Tsao, J., Reboul, J., Vaglio, P., Hill, D. E., Vidal, M., DeLucas, L. J. & Luo, M. (2004). *Genome Res.* **14**, 2102–2110.
Manchee, R. J., Broster, M. G., Melling, J., Henstridge, R. M. & Stagg, A. J. (1981). *Nature (London)*, **294**, 254–255.
Mock, M. & Fouet, A. (2001). *Annu. Rev. Microbiol.* **55**, 647–672.
Ohlsen, K. L., Grimsley, J. K. & Hoch, J. A. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 1756–1760.
Perego, M., Hanstein, C., Welsh, K. M., Djavakhishvili, T., Glaser, P. & Hoch, J. A. (1994). *Cell*, **79**, 1047–1055.
Radnedge, L., Agron, P. G., Hill, K. K., Jackson, P. L., Ticknor, L. O., Keim, P. & Andersen, G. L. (2003). *Appl. Environ. Microbiol.* **69**, 2755–2764.
Ramabhadran, T. V. & Jagger, J. (1976). *Proc. Natl Acad. Sci. USA*, **73**, 59–63.
Read, T. D. *et al.* (2003). *Nature (London)*, **423**, 81–86.
Rodrigues, A. & Hubbard, R. E. (2003). *Brief. Bioinform.* **4**, 150–167.
Romero, P., Obradovic, Z., Li, X., Garner, E., Brown, C. & Dunker, A. K. (2001). *Proteins*, **42**, 38–48.
Siebold, C., Berrow, N., Walter, T. S., Harlos, K., Owens, R. J., Stuart, D. I., Terman, J. R., Kolodkin, A. L., Pasterkamp, R. J. & Jones, E. Y. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 16836–16841.
Studier, F. W. (2005). *Protein Expr. Purif.* **41**, 207–234.
Turnbull, P. C. (2002). C*urr. Top. Microbiol. Immunol.* **271**, 1–19.
Walter, T. S., Diprose, J., Brown, J., Pickford, M., Owens, R. J., Stuart, D. I. & Harlos, K. (2003). *J. Appl. Cryst.* **36**, 308–314.
Walter, T. S., Diprose, J. M., Mayo, C. J., Siebold, C., Pickford, M. G., Carter, L., Sutton, G. C., Berrow, N. S., Brown, J., Berry, I. M., Stewart-Jones, G. B. E., Grimes, J. M., Stammers, D. K., Esnouf, R. M., Jones, E. Y., Owens, R. J., Stuart, D. I. & Harlos, K. (2005). *Acta Cryst.* D**61**, 651–657.
Waterman, D. G., Ortiz-Lombardia, M., Fogg, M. J., Koonin, E. V. & Antson, A. A. (2006). *J. Mol. Biol.* **356**, 97–110.
Yang, Z. R., Thomson, R., McNeil, P. & Esnouf, R. M. (2005). *Bioinformatics*, **21**, 3369–3376.
Zhang, R., Ou, H.-Y. & Zhang, C.-T. (2004). *Nucleic Acids Res.* **32**, D271–D272.