Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Application of the use of high-throughput technologies to the determination of protein structures of bacterial and viral pathogens

M. J. Fogg,[a] P. Alzari,[b] M. Bahar,[c] I. Bertini,[d] J.-M. Betton,[a] W. P. Burmeister,[e] C. Cambillau,[f] B. Canard,[f] M. Carrondo,[g] M. Coll,[h] S. Daenke,[c] O. Dym,[i] M.-P. Egloff,[f] F. J. Enguita,[g] A. Geerlof,[j] A. Haouz,[b] T. A. Jones,[k] Qingjun Ma,[j] S. N. Manicka,[i] M. Migliardi,[d] P. Nordlund,[l]‡ R.J. Owens,[c] Y. Peleg,[i] G. Schneider,[m] R.Schnell,[m] D. I. Stuart,[c] N. Tarbouriech,[e] T. Unge,[k] A. J. Wilkinson,[a] M. Wilmanns,[j] K. S. Wilson,[a] O. Zimhony[i] and J. M. Grimes[c]*

[a]York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5YW, England, [b]Unité de Biochimie Structurale, Institut Pasteur, 25–28 Rue du Dr Roux, 75724 Paris CEDEX 15, France, [c]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Headington, Oxford OX3 7BN, England, [d]CIRMMP, CERM, Via Sacconi 6, 50019 SestoFiorentino, Italy, [e]EMBL-Grenoble c/o ILL, BP181, 6 Rue Jules Horowitz, F-38042 Grenoble CEDEX 9, France, [f]Architecture et Fonction des Macromolécules Biologiques UMR6098, CNRS/Universités de Provence/Université de la Méditerranée Parc Scientifique et Technologique de Luminy, Case 932163, Avenue de Luminy 13288, Marseille CEDEX 09, France, [g]Host–Pathogen Interactions Group, Macromolecular Crystallography Laboratory, ITQB, Ap. 127, 2781-901 Oeiras, Portugal, [h]Departamento de Biología Estructural Instituto de Biología Molecular de Barcelona, Jordi Girona, 1808034 Barcelona, Spain, [i]The Israel Structural Proteomics Center, The Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel, [j]EMBL-Hamburg Outstation, Notkestrasse 85, D-22603 Hamburg, Germany, [k]Department of Molecular Biology of the University Biomedical Center, S-751 24 Uppsala, Sweden, [l]Department of Biochemistry and Biophysics, Stockholm University, S-106 91 Stockholm, Sweden, and [m]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-109 51 Stockholm, Sweden

‡ Present address: Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-109 51 Stockholm, Sweden.

Correspondence e-mail: jonathan@strubi.ox.ac.uk

Received 17 March 2006
Accepted 7 August 2006

The Structural Proteomics In Europe (SPINE) programme is aimed at the development and implementation of high-throughput technologies for the efficient structure determination of proteins of biomedical importance, such as those of bacterial and viral pathogens linked to human health. Despite the challenging nature of some of these targets, 175 novel pathogen protein structures (~220 including complexes) have been determined to date. Here the impact of several technologies on the structural determination of proteins from human pathogens is illustrated with selected examples, including the parallel expression of multiple constructs, the use of standardized refolding protocols and optimized crystallization screens.

1. Introduction and target selection

The aim of the Structural Proteomics In Europe (SPINE) programme was to develop and exploit technological advances in structural biology to tackle difficult problems related to human health and disease. To this end, SPINE workpackage 9 focused on proteins from human pathogens, especially bacterial and viral, whereas work on human targets was addressed by workpackages 10 and 11 (see Banci *et al.*, 2006). This article reports on the progress made in the study of pathogen targets and provides highlights of a number of developments in high-throughput (HTP) technologies which have proved crucial to successful structure determination, illustrated by reference to some of the structures solved. Whilst relevance to human health and disease was the overarching criterion for target selection, a subset of targets was selected explicitly as so-called 'low-hanging fruit'. These, primarily exemplified by the *Bacillus anthracis* proteins (Au *et al.*, 2006), proved particularly valuable in validating HTP pipelines and technologies. The methods used by the various SPINE partners for target selection and other informatics aspects are described in detail in Albeck *et al.* (2006).

1.1. Bacterial targets

The bacterial organisms targeted for structural analysis represent a range of threats to human health, including food poisoning, respiratory diseases and potential bio-terrorism agents, of which all manifest antibiotic resistant strains. The two key organisms were *Mycobacterium tuberculosis* (MTB) and *B. anthracis*, which together accounted for 81% of our bacterial targets and 47% of the bacterial protein three-dimensional structures solved (Table 1). The efforts on MTB involved six partners (Paris, Hamburg, Uppsala, Stockholm, York and the Weizmann) and especially targeted proteins from strain H37Rv. The *B. anthracis* project was a collabora-

Table 1
Bacterial and viral structures solved.

Bacterial targets are shown in normal type, viruses in bold and other organisms in italic. Pathogen species are MT, *M. tuberculosis*; MS, *M. smegmatis*; BA, *B. anthracis*; BS, *B. subtilis*; NM, *N. meningitides*; NG, *N. gonorrhoeae*; CJ, *C. jejuni*; EC, *E. coli*; SP, *Streptomyces purpurascens*; SG, *S. galileus*; SN, *S. nogalater*; SC, *S. calvalugarus*; OT, other bacteria; VV, vaccinia virus; EBV, Epstein–Barr virus; SARS, severe acute respiratory syndrome; MHV, murine hepatitis virus; MVE, Murray Valley encephelitis virus; TEV, tobacco etch virus; BTV, bluetongue virus; BV, Banna virus; ERV, equine rhinitis virus; MV, measles virus.

Species	Protein name	PDB code
MT	Protein phosphatase	
MT	Hypothetical trans factor	2g9w
MT	AhpE	
MT	Hypothetical protein	
MT	Uridylate kinase	
MT	Cyp121	1n4g
MT	Hypothetical protein	1rfe
MT	Thymidylate kinase	1n5j
MT	DapB	1yl6/1yl5/1yl7
MT	LeuB	1w0d
MT	Lipoate protein ligase B	1w66
MT	Putative antiterminator	1s8n
MT	CmtR	2g8p
MT	Pyrophosphatase	1sxv
MT	Sulfite reductase NirA	1zj8/1zj9
MT	Rv1155	1w9a
MT	Rv2945c, lipoprotein	2byo
MT	Hypothetical protein	2fwv
MT	Protein kinase B	1o6y/2fum
MT	Hypothetical protein	2ckd
MT	Hypothetical protein	2bi0
MT	ClpP	2cby/2ce3
MT	Adenylate kinase	2cdn/1p4s
MT	AhpC	2bmx
MT	Rv2740, hydrolase	2bng
MT	PpiA	1w74
MT	Ribose 5-p isomerase B	1usl
MT	Rv1284, carbonic anhydrase	1ylk
MT	Rv3588, carbonic anhydrase	1ym3
MT	Hypothetical protein	2bi0
MT	Hypothetical protein	2c2i
MS	Phosphatidylinositol transferase	2gej/2gek
MS	Protein phosphatase	
BA	Thymidylate kinase BA0027	
BA	Hypothetical protein BA0541	
BA	Phosphoglycerate kinase BA5367	
BA	Spo0E phosphatase BA5174	
BA	Cyclo-ligase BA0296	2btu
BA	Thiolase BA5489	2c5s
BA	UDP-epimerase BA5505	2c2o
BA	Monooxygenase BA2919	
BA	Gly transferase BA1558	
BA	BA0291-s-carboximide	
BA	Ligase BA1563	
BA	Alanine dehydrogenase BA0592	
BA	Cytidine deaminase BA4525	
BA	Nucleoside hydrolase BA2400	2c40
BA	G3PDH 2 BA5369	
BA	G3PDH 1 BA4827	
BA	P-kinase BA3382	
BA	P-kinase BA4843	
BA	Cycloligase BA4489	
BA	Translation factor P BA4421	
BA	Uridylate kinase BA1797	
BA	Naphthoate synthase BA5109	
BA	Docking protein ftsy BA3985	
BA	Hydratase/isomerase BA3583	
BA	Alanine racemase BA0252	
BA	LOLS protein BA4318	
BA	Thi1 BA4899	
BA	Met-aminopeptidase BA0132	

Table 1 (continued)

Species	Protein name	PDB code
BA	3-Oxo-acyl reductase BA3989	
BA	Ribulose-p-epimerase BA3998	
BA	Histidyl-tRNA synthetase BA4633	
BA	Enolase BA5364	
BA	Arginase BA0154	
BA	Hydratase/isomerise BA2356	
BA	YisI, BA1655	2bzb
BA	Dihydrodipicolinate synthase BA3935	1xky/1xly
BA	PN phosphorylase BA1483	1xe3
BA	pure BA0288	1xmp
BA	GuaC BA5705	1ypf
BA	Ferrochelatase BA1071	2c8j
BA	Endonuclease IV BA4508	1xp3
BA	Superoxide dismutase BA4499	1xre
BA	Superoxide dismutase BA5696	1xuq
BS	Forespore regulator Bsu2771	2bw2
BS	dUTPase YncF Bsu1767	
BS	Transcription regulator	2b182bo1
BS	CsrA	1t30
BS	dUTPase YosS Bsu2001	
BS	PhoP	1mvo
BS	P104H BsSOD	1xtl
BS	Y88H-P104H BsSOD	1xtm
BS	AppA	1xoc
BS	Sco1	1on4
BS	S46V CopAa	1oq3
BS	S46VCopAab	1p6t
BS	SOD-like protein	1rp6
BS	SOD-like protein	1s4i
NM	LysR transcription factor	
NM	MarR transcription factor	
NM	Lrp transcription factor	
NM	Yjgf family regulator	
NM	Pii GlnB protein	2gw8
NM	MarR	
NM	IIA^{ntr} protein	2aoj
NG	Specific regulator	
CJ	SurE survival protein	
CJ	Periplasmic binding protein	
CJ	Solute-binding protein	
CJ	Cysteine-binding protein	1xt8
CJ	dUTPase Cj1451	1w2y
EC	Carboxypeptidase Dcm	1y79
EC	ATP NAD kinase	
EC	Biosynthesis-like protein	
EC	tRNA pseudouridine syn	1szw
EC	Diacylg-kinase catalytic	1bon
EC	Type III CoA transferase	1xa3
EC	Protein ribD	
EC	Hyp-protein yebU	
EC	UPF0010 protein	
EC	CutA1	1naq
SP	Aclacinomycin m-esterase	1q0z
SP	Aclacinomycin hydroxylase	1qzz
SP	C-4-*O*-methyltransferase	1tw2/1tw3
SG	Anthracycline 1-hydroxylase	
SN	NA methyl-ester-cyclase	1sjw
SN	Snoal2 1-hydroxylase	
SC	BLIP_Mut1	
SC	BLIP_TEM_2	1xxm
SC	BLIP_TEM_3	1s0w
OT	P275D-TbADH	
OT	Tb_Eh_ADH	
OT	Cb_Tb_ADH	
OT	Tb_Cb_ADH	
OT	PaADH	
OT	D275P_EhADH	
OT	DiVK	1mb3
OT	CopC	1ot4
OT	1,3-PD-dehydrogenase	
OT	Carbonate dehydratase	
OT	Periplasmic binding protein	

Table 1 (continued)

Species	Protein name	PDB code
OT	Alcohol dehydrogenase	2b83
OT	P-carrier protein hpr	1ka5
OT	P-carrier protein hpr	1txe
OT	Atx1-apo	1sb6
OT	cadA ATPase	2aj0/2aj1
OT	P275D-TbADH	
OT	CheY2	1p6q
OT	tRNA synthetase homolog	
OT	CR1885	1x9l
OT	DiVK	1*mb*3
OT	ctag	1so9
VV	**B14R**	
VV	**N1L**	
VV	**CmrE**	
VV	**A41L**	
EBV	**U-DNA glycosylase**	
EBV	**BKRF3**	
EBV	**BLLF3**	**2bsy**
EBV	**BVRF2**	**1o6e**
EBV	**BARF1**	**2ch8**
SARS	**Macro domain**	**2fav**
SARS	**Orf 9 b**	**2cme**
SARS	**Entire NSP9**	**1qz8**
SARS	**Nsp9**	**1uw7**
MHV	**Nsp9**	
MVE	**Methyltransferase**	
TEV	**NIa protease**	
PM2	**Spike protein P1**	
Phi-6	**P2 polymerase**	
Phi-13	**P4 translocase**	
BTV	**VP4 guanylyltransferase**	
BTV	**Nonstructural ns2**	
BV	**VP9 capsid protien**	
ERV	**ERAV capsid**	
MV	**Domain of P**	**2bet**
Bil170	**Bil170 RBP**	
p2	**P2 RBP/complex**	**2bsd/2bse**
Tick	*TdPI*	
Tick	*TdPI/complex*	
Tick	*dS salivary peptide*	
Prot	*Carbonic anhydrase*	*1y7w*

tive effort between the Oxford and York partners and a fuller assessment of this is given in an accompanying paper (Au *et al.*, 2006). In addition to these, smaller numbers of targets were selected from *Campylobacter jejuni* (York, Lisbon), two members of the *Neisseriae* family (*N. meningitidis* and *N. gonorrhoeae*; Oxford), *Escherichia coli* (*e.g.* 0157:H7; Stockholm, Weizmann, Florence), *Klebsiella pneumoniae* (Lisbon) and *Streptomycetes* (Stockholm) and *Shigella flexnin*, *Salmonella entrica typhimurium* and *Streptococcus pyogenes* (Weizmann).

1.1.1. *M. tuberculosis*. A number of slow-growing mycobacterial species have evolved into highly successful human pathogens and are responsible for diseases such as tuberculosis (TB; *M. tuberculosis* and *M. africanum*) and leprosy (*M. leprae*). The World Health Organization estimates that one-third of the human population is currently infected and that one person dies roughly every 18 s (http://www.who.int) from mycobacterial diseases. There are now forms of *M. tuberculosis* (MDR-TB; Espinal, 2003) resistant to the two most powerful anti-TB drugs (isoniazid and rifampicin) and there is a pressing need for new drugs effective against persistent TB infection and MDR-TB. Closely related forms of mycobacteria are pathogenic to other mammalian species, *e.g.* *M. bovis*. The MTB target-selection process in Paris is illustrative of those used for this organism in SPINE and was based on comparative analyses of the complete genome sequences for two MTB strains with other bacteria and mycobacterial strains, such as *M. bovis*, *M. bovis* BCG and *M. leprae*, which provided detailed information about every gene in MTB. Over 300 proteins were identified that were only found in mycobacteria or actinomycetes, but were conserved in the degraded genome of *M. leprae* and therefore presumed to be important for bacterial viability and hence potential drug targets (Albeck *et al.*, 2006). In addition, the Paris laboratory targeted proteins for which structures would provide insights into MTB biology. Funding for the MTB project was far from exclusive to SPINE and came from several other sources, including other EC grants.

1.1.2. *B. anthracis*. York and Oxford each selected initial sets of 48 targets from *B. anthracis*, with the dual aim of determining a reasonable number of structures and, in the process, helping to refine their protein-production pipelines (Au *et al.*, 2006). To be considered, the protein had, in general, to be reasonably small (<50 kDa), lacking in regions predicted to be transmembrane moieties, signal peptides or disordered and usually amenable to structure solution by molecular replacement. York targets included proteins involved in nucleotide metabolism and sporulation, whilst Oxford targets included members of protein families well conserved across a range of pathogenic bacteria (Au *et al.*, 2006). An additional SPINE collaboration between York and Utrecht resulted in two NMR structures of sporulation-associated aspartic acid phosphate phosphatases, neither of which could be crystallized (AB *et al.*, 2006), and extended the targets to the BofC (bypass of forespore C) protein from *B. subtilis*, an intercompartmental signalling factor expressed in the forespore (Patterson *et al.*, 2005), also found in *B. anthracis* (BA4653).

1.1.3. *Neisseria*. Many species of this Gram-negative β-proteobacteria are found only in humans. Two members of the *Neisseriae*, *N. meningitidis* and *N. gonorrhoeae*, have evolved into highly successful pathogens responsible for bacterial meningitis and gonorrhoea, respectively. Although the regulatory systems of all Gram-negative bacteria share features in common with *E. coli*, *Neisseria* species display several features which are unique. These include a restricted set of sigma factors, a relatively small number of transcriptional regulators and a large number of phase variable genes. Oxford focused on the complete repertoire of transcription regulators in both *N. meningitidis* and *N. gonorrhoeae*, with a view to defining the structure–function relationships of these proteins and to provide key reagents for experimental studies of regulation. The four available complete genome sequences (Parkhill *et al.*, 2000; Tettelin *et al.*, 2000) of *Neisseria* were assessed for homologues of DNA-binding proteins with potential regulatory functions. Representative orthologues of each regulator in each genome were identified, with the genes selected in order of preference from *N. meningitidis* strain MC58, *N. gonorrhoeae* strain FA1090, *N. meningitidis* strain

Z2491 and *N. meningitidis* strain FAM18, to provide a cohort of 62 DNA-binding and associated signal-transduction proteins. In the vast majority of cases, the entire protein was targeted for analysis.

1.1.4. A pan-bacterial attack on copper-binding proteins. To address the role of copper in cellular processes, the Florence laboratory targeted copper-binding proteins. Bacterial and eukaryotic genomes were compared to identify prokaryotic orthologues of proteins in yeast and humans. The aim was to solve high-resolution NMR structures that could be used as models for human proteins and to identify differences exploitable for potential antibacterial drugs. Additionally, target selection was limited to proteins of fewer than 200 residues, as they are more tractable for NMR (Banci & Rosato, 2003). Structures were solved from a range of bacterial species including *B. subtilis*, *E. coli* and *Pseudomonas aeruginosa* (AB *et al.*, 2006).

1.2. Viruses

A number of different viruses were targeted, ranging from large double-stranded DNA (dsDNA) viruses such as Epstein–Barr virus (EBV; Grenoble) and vaccinia virus (VACV; Oxford), to single-stranded RNA viruses such as the newly emergent human pathogen, the severe acute respiratory syndrome coronavirus (SARS-CoV; Marseille and Oxford). Finally, some viruses of less direct relevance to human health were investigated as model systems, for instance bacteriophage P2 (Marseille).

1.2.1. Poxviruses. This group of very large viruses includes variola virus, the aetiological agent of smallpox, and vaccinia virus (VACV), used as the smallpox vaccine. Their genomes each contain ∼200 open reading frames (ORFs) and the sequences of more than 20 poxviruses reveal considerable sequence variability towards each end of the genome. These regions code for genes that affect virus virulence, host-cell susceptibility and the host response to infection (Gubser *et al.*, 2004), enabling the virus to suppress the immune response (Smith *et al.*, 1997, 1999). In VACV half of the genes are non-essential for replication and a significant number are immunomodulators; nine of which were targeted by Oxford as being reasonably small but of significant potential biological interest.

1.2.2. Herpesviruses. Members of this family of dsDNA viruses have large genomes (∼100 ORFs) and are responsible for a diverse set of human diseases. Examples include herpes simplex virus (HSV; causing genital herpes and herpes encephalitis), EBV (causing Burkitts lymphoma and glandular fever), varicella zoster virus (VZV; causing chickenpox and shingles), human herpesvirus-8 (HHV-8; implicated in all forms of Karposi sarcomas, commonly seen in AIDS patients) and human cytomegalovirus (CMV; causing AIDS-related retinitis and pneumonia). No effective treatments are generally available, with the exception of the acyclovir compounds active against HSV. Thus, herpesviruses are of immediate interest to the biomedical community and in particular to the pharmaceutical industry. Barcelona and Oxford have interests in the molecular mechanisms of genome packaging and targeted a number of proteins that form the complex that packages the dsDNA genome of herpesviruses. This work demonstrated the inadequacy of the current *E. coli* protein-production pipelines for such difficult proteins (five out of six of the components are greater than 70 kDa in size and all are part of complex robust macromolecular assemblies), since all the expressed proteins were insoluble. Soluble proteins have now been produced in insect cells for some of these targets, but no structures have been determined and this work will not be considered further here. In contrast, the Grenoble group targeted EBV proteins with known enzymatic activity and ranked them based on a number of properties, prioritizing those that were small and predicted to be stable with a high secondary-structure content. Surface glycoproteins and those proteins forming multi-component complexes were explicitly excluded. Application of these selection criteria led to the structure determination of four proteins from EBV (Tarbouriech *et al.*, 2006).

1.2.3. Coronaviruses. SARS emerged as a new human disease in southern China in late 2002. The first manifestation of infection is a febrile illness, with respiratory symptoms, headaches and myalgia, followed by progression to acute respiratory distress and progressive respiratory failure (Peiris *et al.*, 2003). The disease is caused by a coronavirus (Kuiken *et al.*, 2003), one of a group of enveloped positive-strand RNA viruses commonly associated with enteric and respiratory disease (Ziebuhr & Siddell, 2002). The unusual severity of SARS-CoV infection probably reflects the introduction of an animal coronavirus into a susceptible human population. In the first outbreak at least 8000 people were infected and there were more than 750 fatalities (Donnelly *et al.*, 2003). The SARS-CoV genome is composed of at least 14 functional ORFs that encode three classes of proteins: structural (S, M, E

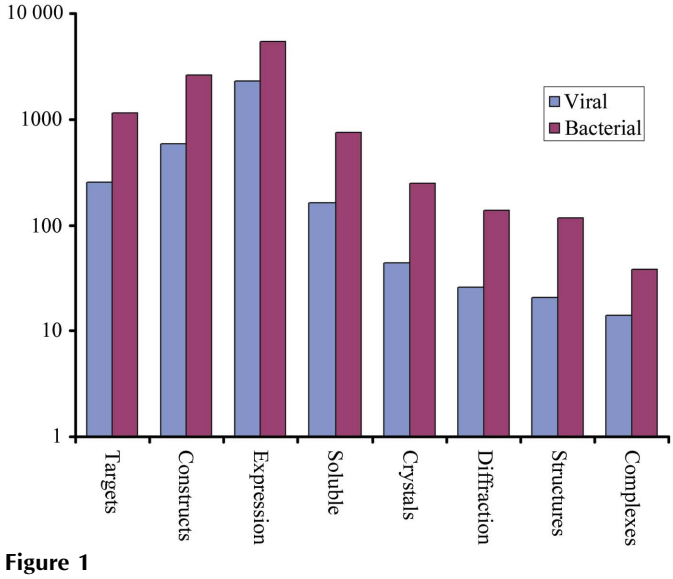

Figure 1
SPINE bacterial and viral target scoreboards. A snapshot (September 2005) of the pipeline for both bacterial and viral targets, drawn as a bar chart, with bacterial targets in purple and viral in blue. The vertical axis is represented on a log scale.

and N), non-structural involved in viral RNA synthesis (nsp or replicase) and proteins thought to be non-essential for replication in tissue culture but that clearly provide a selective advantage *in vivo* (the nspX or accessory proteins; Marra *et al.*, 2003; Rota *et al.*, 2003). A number of proteins, primarily from the large replicase polyproteins ORF1a and ORF1ab of SARS-CoV, have been targeted by Oxford and Marseille to test the efficacy of a focused structural proteomics approach against a newly emerging human pathogen. Proteins were excluded on the basis of certain criteria, such as size, glycosylation and the presence of transmembrane regions.

2. Scoreboard for bacterial and viral proteins

As with all large-scale projects, the 'Scoreboard' is only a snapshot of target progress through the various pipeline stages and is therefore not a definitive assessment. It does, however, allow an assessment of whether there are particular bottlenecks that need addressing. The numbers for this section are from a comprehensive snapshot of the project taken in September 2005 (overall summary presented in Fig. 1), whereas the detailed numbers for individual projects are for early 2006 (Table 1). A comparison of the number of expression trials with the number of crystals or good HSQC data sets for the entire SPINE project shows that around 10% of constructs led to the production of a protein suitable for structural analysis. This attrition arises from approximately equal losses at the expression stage and crystallization or sample preparation stages (for NMR).

Most of the proteins were expressed in prokaryotic systems; however, there is evidence, at least for human and viral proteins, that significant gains are possible by the use of eukaryotic expression systems, thus a global analysis of results from SPINE partners shows that the success rate for mammalian cell expression is more than twice that for soluble expression in *E. coli* (75% *versus* 30%), with insect-cell expression giving an intermediate success rate (44%). Further development and uptake of these methodologies (see Aricescu, Assenberg *et al.*, 2006; Aricescu, Lu *et al.*, 2006) is likely to have a major impact on the study of difficult viral targets, as indicated by preliminary results for EBV proteins, where insect-cell expression yielded soluble protein for 50% of the proteins tested (Tarbouriech *et al.*, 2006)

2.1. Bacterial targets

The production of soluble protein for NMR and crystallization studies is an absolute requirement, yet it remained a significant bottleneck for the SPINE bacterial targets. Bacterial expression statistics from all contributing SPINE partners are illustrated in Fig. 1. Over 1000 proteins were targeted and on average two or three constructs were produced for each target and each construct was subjected to two expression trials, with 29% of constructs producing soluble protein. A more detailed analysis reveals significant variability, with 77% of the first cohort of *B. anthracis* constructs expressed in soluble form, but only 33% for *MTB* and 15% for other bacteria. As of September 2005, almost 35% of these soluble proteins had produced protein crystals or data suitable for NMR; these had resulted in structure solutions for slightly more than 10% of the original 1127 targets (Fig. 1). However, target groups which were carefully selected for amenability for expression and structure determination, such as the *B. anthracis* cohort, fared better, with success rates in crystallization approaching 60% (Au *et al.*, 2006).

2.2. Viral targets

The SPINE viral scoreboard (Fig. 1) suggests that although viral proteins have a lower success rate for producing soluble proteins than *B. anthracis* proteins, they are comparable to more difficult bacterial and human proteins (~27% of constructs resulted in soluble expression at a level sufficient for production and purification). However, the attrition rate from protein purification to structure determination is higher, with only about 30 crystal structures having been determined to date. It appears that this is in part a consequence of the difficulty of growing high-quality crystals. Thus, 34% of soluble proteins crystallized compared with an equivalent rate of 58% for the York/Oxford *B. anthracis* targets; furthermore, many crystals were not of diffraction quality. However, as for the bacterial targets, patterns of soluble protein expression vary between viral systems. For some groups of viruses, such as the poxviruses, the success rate for production in *E. coli* was reasonable, as discussed below.

3. Pipeline technologies and case studies of their impact

The protein-production technologies developed and used by SPINE partners are covered in detail by Alzari *et al.* (2006) and Aricescu, Assenberg *et al.* (2006) and aspects of work on specific pathogen proteins are presented by Au *et al.* (2006) and Tarbouriech *et al.* (2006). Target genes were often amplified, cloned and screened for expression in parallel (*e.g.* 48 or 96 at a time). Ligation-independent cloning strategies were widely (but not exclusively) used and the use of N- and C-terminal His_6 tags was almost universal. The automation of many procedures, including small-scale expression screening, was achieved using liquid-handling robots, with multi-channel pipettes providing an inexpensive option for less automated pipelines (expression-screening strategies are compared by Berrow *et al.*, 2006). Biophysical characterization methodologies usually included dynamic light scattering (DLS) and mass spectrometry (ESI–MS) (see Geerlof *et al.*, 2006). Protein crystallization on the nanolitre scale was used by many of the partners (Berry *et al.*, 2006). Here we illustrate the impact of a number of particular SPINE-based technologies on the structure determination of pathogen targets, namely (i) multiple-construct design for a single target, (ii) optimization of protein solubility, (iii) eukaryotic expression systems, (iv) standardized refolding protocols, (v) surface engineering and (vi) other biophysical characterization.

3.1. Multiple-construct design

The use of multiple constructs to define the correct N- and C-termini of a target domain has had a crucial impact on the successful structural analysis of both pathogen and human targets (Banci *et al.*, 2006; Siebold *et al.*, 2005). An example of this in the context of viral proteins is the N-terminal domain of a viral methyltransferase, NS5, of Murray Valley encephalitis virus (MVE), which was studied in Oxford. The domain boundary was already reasonably well defined from the crystal structure of the related methyl-transferase domain from dengue virus (Egloff *et al.*, 2002) and 18 constructs, comprising of nine pairs of N- and C-terminal His_6-tagged sequences, were designed that varied the C-terminal end of the protein domain. On the basis of automated small-scale expression screening, six constructs were selected for production. DLS indicated that all protein samples were monodisperse and ESI–MS confirmed the molecular weights for the samples. The His tags were cleaved (using the rhinovirus 3C cleavage site for N-terminal His tags or by carboxypeptidase A treatment to cleave back to a Lys for C-terminal His tags) and these samples entered nanolitre-scale crystallization trials using the standard Oxford set of 672 screens (Walter *et al.*, 2005). Since none of these screens gave crystals, the proteins were subjected to lysine methylation (as described below), resulting in crystals and a structure for one construct.

3.2. Solubility and functional optimization

Efforts to improve the solubility of MTB proteins illustrate how the SPINE partners have tackled the problem of insoluble expression. MTB proteins are notoriously insoluble or provide poor yields when expressed heterologously in *E. coli* systems (Bellinzoni & Riccardi, 2003; Alzari *et al.*, 2006), a possible consequence of the high GC content of MTB genes (65–70%) and unique codon preferences (de Miranda *et al.*, 2000). To address such problems, Hamburg contributed to the development of an MTB expression system based on the faster growing close relative *M. smegmatis* (Daugelat *et al.*, 2003). Proteins were expressed under close to physiological conditions in order to ensure the correct post-translational modifications such as glycosylation and methylation. The published protocol, using an inducible acetamidase promoter, was modified to be compatible with the EMBL pETM vector system (Geerlof *et al.*, unpublished work). Using this system, the MTB LipB enzyme was expressed with a covalently bound ligand, which resulted in an X-ray structure at 1.08 Å resolution (Ma *et al.*, 2006). When expressed in *E. coli*, a mixture of native and ligand-bound enzyme was produced which failed to crystallize. *M. smegmatis* expression was also used by the Weizmann for the MTB multidomain eukaryotic like fatty-acid synthase I (FASI; Zimhony *et al.*, 2004) after *E. coli* expression had yielded a soluble but non-functional FASI. A recombinant *M. smegmatis* strain was constructed by deleting the native *fas1* gene and replacing it with the MTB (H37Rv) *fas1* gene using a site-specific integrating cosmid carrying the MTB *fas1* gene and its 5′ and 3′ flanking regions. This produced a faster growing non-pathogenic *M. smegmatis* (mc2155) derivative expressing MTB FASI.

In Paris, two strategies were adopted to enhance protein solubility. Firstly, parallel cloning of orthologous mycobacterial genes from different species (*i.e. M. tuberculosis*, *M. leprae*, *M. smegmatis* and *M. bovis*) was used to identify the most soluble candidate. Secondly, cell-free protein expression (Roche RTS) allowed the rapid evaluation of single gene constructs. In the latter strategy, N-terminal codons were optimized to identify the best silent mutation(s) for maximum expression (Betton *et al.*, 2004). Such cell-free systems provide a fast standardized method for screening potentially toxic protein expression without the inherent risks of using live cells. Finally, in Stockholm, the MTB NirA protein could only be obtained in a soluble form in *E. coli* following cloning and co-expression with the *cysG* gene from *Salmonella typhimurium*, a gene that codes for an enzyme catalysing three steps in the biosynthesis of sirohaem, a cofactor of NirA.

3.3. Baculoviral expression

Baculovirus-driven expression in insect cells has been used for a number of viral proteins that were resistant to soluble expression in *E. coli*. Although screening for expression in such systems is at an early stage of development, the technology offers a viable alternative to bacterial expression for 'high-value' targets, as mentioned above for the EBV work. An example from Oxford is the work on the capping enzyme of bluetongue virus VP4, which, owing to problems of toxicity in bacteria, was expressed in insect cells. This protein was problematic during both purification and crystallization, requiring the presence of high salt concentrations throughout. In addition, production of SeMet-derivatized protein proved difficult, prompting development of improved protocols for labelling in insect cells (Sutton *et al.*, unpublished work; Aricescu, Assenberg *et al.*, 2006).

3.4. Standardized refolding

Refolding from *E. coli* inclusion bodies was applied in a number of viral projects, with varying degrees of success. The extracellular immunomodulators of VACV studied in Oxford provide a good example. These targets may be grouped according to how they act: (i) secreted proteins that bind to host factors that regulate complement, interferon (IFN), chemokines, cytokines and inflammation (Smith & Alcami, 2000), (ii) intracellular proteins that modulate signalling pathways, apoptosis or the antiviral action of IFN (Alcami *et al.*, 1999; Tortorella *et al.*, 2000) and (iii) others that are present on the infected cell's surface and modulate the interaction of the infected cell with host factors and other cells. Nine secreted immunomodulators of VACV were targeted for expression, of which four were successfully refolded from inclusion bodies (for the protocol, see Alzari *et al.*, 2006), resulting in the structure determination of two targets, one of which, A41L, is highlighted below.

3.5. Surface engineering as a tool in structural genomics

As noted above, a major bottleneck, particularly for the viral targets, was the growth of crystals that were sufficiently well ordered for structure solution. The crystallization of proteins is dependent on their surface properties and it is well established that flexible side chains (such as lysine residues) on the surface of proteins reduce the likelihood of successful crystallization (see, for example, Rayment, 1997). A straightforward method for changing the surface properties ('surface engineering') of proteins is the reductive methylation of lysine residues and this has been shown in some cases to yield samples that are more amenable for growth of high-quality crystals (Kobayashi *et al.*, 1999; Kurinov *et al.*, 2000; Schubot & Waugh, 2004). To test the effectiveness of this method, Oxford developed a simple, cheap and robust protocol and carried out a systematic study of eukaryotic, bacterial and viral targets. Ten proteins which had previously proved refractory to structural analysis were successfully methylated and entered into crystallization trials (Walter *et al.*, 2006, in the press). To date, this has led to three novel structures contributed to workpackage 9, one being the MVE methyltransferase domain discussed above.

3.6. Other biophysical characterization

The greatly enhanced use of MS for routine protein characterization was a significant outcome of SPINE and proved of value in numerous ways; for example, ESI–MS aided a revised functional assignment of a *C. jejuni* extracytoplasmic solute receptor (Cj0982), putatively annotated as a glutamine-binding protein from its amino-acid sequence. The ESI–MS spectrum of purified Cj0982 identified a peak with a mass of 125 greater than the mass of the protein alone. Following crystallization and X-ray structure solution, a bound L-cysteine ligand (mass 125) was identified (Müller *et al.*, 2005). Standard procedures have been developed in Oxford to use ESI–MS to determine levels of post-translational modification such as glycosylation (Geerlof *et al.*, 2006). NMR target proteins were checked for correct folding using ^{1}H–^{15}N heteronuclear single-quantum correlation (HSQC) spectroscopy and metal-binding properties checked by atomic absorption, using extended X-ray absorption fine structure (EXAFS) spectroscopy, UV–Vis spectroscopy and (where appropriate) electron paramagnetic resonance (EPR) spectroscopy. In several laboratories (Marseille, Oxford, Stockholm), Thermofluor analysis is routinely used to monitor the temperature-dependence of protein unfolding (Geerlof *et al.*, 2006) to define compounds to guide the optimization of crystallization conditions by stabilizing the protein.

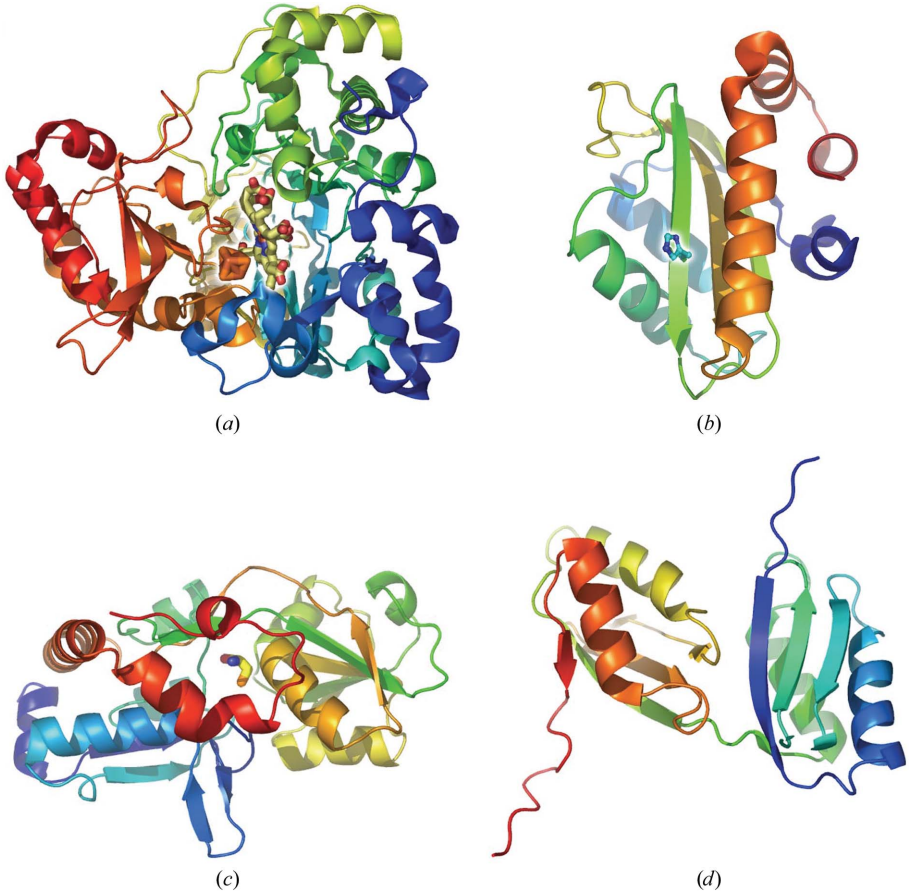

Figure 2
Structures of representative bacterial targets described in §4. All cartoons are drawn blue to red from the N- to the C-terminus. (*a*) Schematic view of the structure of sulfite reductase, NirA, from *M. tuberculosis*. The [Fe_4–S_4] cluster and siroheme cofactor are shown in ball-and-stick representation (Schnell *et al.*, 2005). (*b*) Schematic view of the structure of *Neisseria* IIAntr (NMB 0736) with the phosphoryl acceptor histidine residue (His67) indicated (Ren *et al.*, 2005). (*c*) Schematic view of the structure of *C. jejuni* Cj0982 with bound cysteine ligand shown in ball-and-stick representation (Müller *et al.*, 2005). (*d*) Schematic view of the NMR structure of *B. subtilis* CopA N-terminal domains (Banci *et al.*, 2003*a*,*b*).

4. Selected highlights

Here, we present examples chosen to reflect different aspects of the broad target areas: NirA from MTB and the dUTPase of EBV are potential drug targets, while work on *Neisseria* and copper-binding proteins attempts to use structure to further illuminate function and the work on *C. jejuni*, SARS-CoV and VACV shows how structure can shed light on the possible roles for proteins of unknown function. Finally, the work on bacteriophage P2 shows how a test case, used to evaluate HTP procedures, can provide biologically important data.

4.1. *M. tuberculosis*

Overall, 31 MTB structures have been solved to date within SPINE and here we provide a single example. One

of the MTB genes upregulated in its persistent state is the essential gene *NirA* (Rv2391), suggesting that it is a potential target for the development of anti-tuberculosis agents. The product of this gene, NirA protein, has been targeted by the Stockholm group. The amino-acid sequence of NirA is homologous to ferredoxin-dependent sulfite reductases and the purified recombinant enzyme demonstrates a characteristic [Fe_4–S_4] absorption spectrum. The structure of the enzyme was determined using X-ray crystallography (Fig. 2; Schnell *et al.*, 2005). For this and other difficult projects (including many MTB proteins), it was necessary to employ a hand-crafted approach, exemplified by the requirement to use streak-seeding to improve crystals and by the requirement for co-expression noted above (§3.2). At the active site, the side chains of Tyr69 and Cys161, located in close proximity to the sirohaem cofactor, form an unusual covalent bond. Mutagenesis of these residues decreased the catalytic activity of the enzyme, with Y69A and C161S having the most deleterious effect, suggesting that the covalent bond is important but not essential for enzyme activity. These residues are part of a sequence fingerprint which distinguishes ferredoxin-dependent sulfite from nitrite reductases. This is the first three-dimensional structure obtained of a ferredoxin-dependent sulfite/nitrite reductase.

4.2. *Neisseria*

Genes for the cohort of 62 DNA-binding and associated signal-transduction protein were amplified, cloned and screened for expression in parallel and the proteins were expressed, purified and crystallized in parallel using the standard OPPF pipeline. So far, the structures of seven have been solved, including the first bacterial leucine response regulator (Lrp) protein (NMB1650), a global regulator of amino-acid metabolism (Calvo & Matthews, 1994) and a MarR family regulator (NMB1585). In *E. coli*, Lrp is a global regulator of amino-acid metabolism (Calvo & Matthews, 1994), whereas MarR is a specific regulator of the multiple antibiotic resistance operon (Alekshun & Levy, 1997). Studies are in progress to relate the structures of the *Neisseria* transcription factors to their function in this organism and Thermofluor analysis has been used to identify the amino-acid cofactor of Lrp (Nichols *et al.*, manuscript in preparation). Amongst the signal transduction proteins, it has been shown that the structure of the *N. meningitidis* nitrogen regulatory protein IIAntr (NMB0736; Fig. 2) confirms its assignment as a functional homologue of the IIAntr proteins found in a range of other Gram-negative bacteria (Ren *et al.*, 2005).

4.3. *C. jejuni*

The Gram-negative pathogen *C. jejuni* is the leading cause of gastroenteritis in humans and is responsible for ∼30% of Guillain–Barré syndrome cases, a debilitating polio-like autoimmune disease. *C. jejuni* colonizes the intestinal tract of many animals (Svedhem & Kaijser, 1981) and is commensal in poultry, cattle and swine (Harris *et al.*, 1986; Kazwala *et al.*, 1990). The most promising route to reducing infections is to reduce the incidence in poultry *via* vaccination. The *C. jejuni* genome (Parkhill *et al.*, 2000) harbours several genes encoding highly antigenic proteins, three of which are the periplasmic binding component of an ABC amino-acid transport system

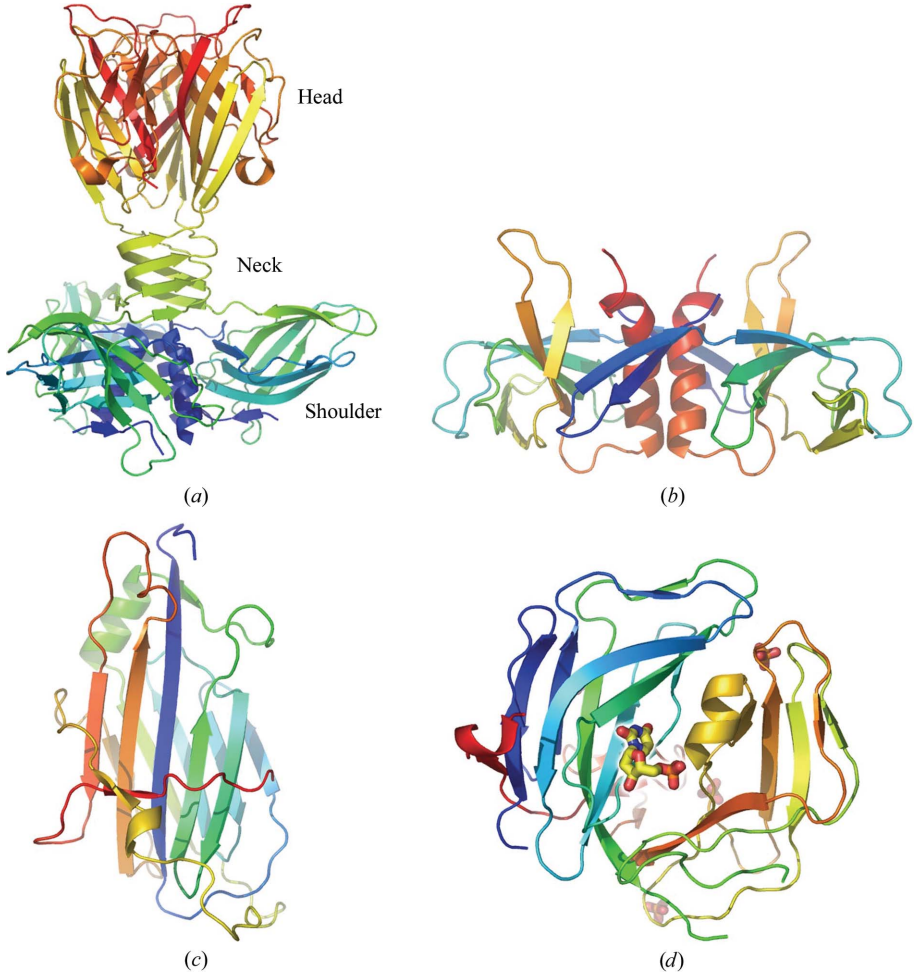

Figure 3
Structures of representative viral targets described in §4. All cartoons are drawn with each monomer coloured from blue to red from the N- to the C-terminus. (*a*) The trimeric receptor-binding protein from lactococcal bacteriophage P2. The threefold axis is vertical and the separate architectural units of the molecule are labelled according to Spinelli *et al.* (2006). (*b*) The dimeric molecule of SAR-CoV nsp9, viewed orthogonal to the molecular twofold axis. (*c*) Ribbon diagram of A41L of VACV (M. Bahar, unpublished work). (*d*) Ribbon representation of EBV dUTPase in complex with an atomic representation of dUMP (C, O, N and P atoms drawn in yellow, red, blue and orange, respectively).

and include Cj0982, a major surface antigen and vaccine candidate (Wyszynska *et al.*, 2004). Such amino-acid uptake systems are crucial to *C. jejuni* since it relies on amino acids as a carbon source (its incomplete glycolytic pathway renders it unable to use sugars; Kelly, 2001). York determined the crystal structure of Cj0982, revealing that the ligand-binding pocket contained a cysteine (Fig. 2), which was confirmed by ESI–MS and tyrosine-fluorescence spectroscopy, suggesting that the protein belongs to a cysteine-binding ABC-transport system (Müller *et al.*, 2005). This is the first structure of a cysteine-transport protein. Crystallization experiments were initially carried out at the Oxford Protein Production Facility and were refined using an in-house Mosquito nanolitre-dispensing robot. Five *C. jejuni* structures have been solved as part of the SPINE project.

4.4. Copper-binding proteins

The Florence group have solved 20 novel structures of copper-binding proteins by NMR. The copper homeostasis systems they have characterized by NMR include the P-type ATPase CopA from *B. subtilis* (Banci *et al.*, 2002), a non-pathogenic close relative of *B. anthracis* that also contains a close CopA orthologue (BA3859). An unusual feature of these P-type ATPases is that the cytoplasmic N-terminus has one or more domains, depending on the complexity of the organism, each containing a metal-binding motif. *B. subtilis* CopA has two domains (CopAa and CopAb) with similar amino-acid sequences, whereas the copper ATPases of other bacteria have only one domain. NMR analysis revealed one *B. subtilis* domain (CopAa) to be largely unstructured, suggesting that it may be unfolded *in vivo* and not functional (Banci *et al.*, 2002). Sequence comparison of the two domains and orthologous proteins suggested that Ser46 of CopAa may destabilize the hydrophobic core of the domain, as hydrophobic residues (Val and Ala) occur frequently at this position. Indeed, the single mutation S46V of CopAa produced a completely folded domain. The structure of the CopA N-terminal domain was solved by NMR (Fig. 2) and the interaction with its copper chaperone partner CopZ (from the same operon) was also characterized (Banci *et al.*, 2003*a*,*b*). It is hypothesized that *in vivo* the folded functional domain is favoured over an unfolded state, in a manner analogous to the way that the N-terminus of the Wilson's disease protein interacts with its ATP-binding domain (Banci *et al.*, 2003*a*).

4.5. Receptor-binding protein of bacteriophage P2

Marseille have investigated, as a test case, the mechanism of cell entry by bacteriophage P2. This tailed phage infects *Lactococcus lactis*, a Gram-positive bacterium used extensively for the manufacture of fermented milk products. Infection of *L. lactis* by tailed phages leads to serious financial losses. Bacteriophage P2 infects specific *L. lactis* strains by using a receptor-binding protein (RBP) located at the tip of its non-contractile tail. As with the NirA protein, described above, crystals suitable for X-ray diffraction experiments could only be obtained by seeding. Bacteriophage P2 RBP is a homotrimeric protein, each subunit of which comprises three domains (Fig. 3): the shoulders (a β-sandwich attached to the phage), the neck (an interlaced β-prism) and the receptor-recognition head (a seven-stranded β-barrel; Spinelli *et al.*, 2006). The complex of RBP with a neutralizing llama VHH5 domain allowed identification of the area on RBP that attaches to the bacterial receptor (Spinelli *et al.*, 2006), where it is able to bind various saccharides (Tremblay *et al.*, 2006). The structural similarity between the recognition-head domain of bacteriophage P2 and those of adenoviruses or reoviruses, which invade mammalian cells, suggests that these viruses, despite being evolutionarily distant and having different chemical genomic composition (DNA *versus* RNA), may have a common ancestral gene.

4.6. SARS-CoV nsp9

The SARS-CoV replicase gene encodes multiple enzymatic functions (Snijder *et al.*, 2003). These include an RNA-dependent RNA polymerase activity (RdRp, nsp12), a 3C-like serine proteinase activity (3CLpro, nsp5, also known as the main proteinase, Mpro), a papain-like proteinase activity (PL2pro, nsp3) and a superfamily 1-like helicase activity (HEL1, nsp13). These types of proteins are common to the replicative machinery of many positive-strand RNA viruses. In addition, the replicase gene encodes proteins that have domains likely to possess enzymatic activities associated with RNA modification. Marseille and Oxford have studied SARS proteins and have to date solved four separate structures using the standard protein-production and crystallization pipelines described by Alzari *et al.* (2006) and Berry *et al.* (2006). One of these, the replicase protein nsp9, was solved by both partners (Egloff *et al.*, 2004; Sutton *et al.*, 2004) and we discuss this structure below.

Nsp9, encoded by ORF1a, has no designated function but is most likely involved in viral RNA synthesis. The protein comprises a single β-barrel with a fold previously unseen in single-domain proteins (Fig. 3). The topology superficially resembles an OB-fold with a C-terminal extension and is related to both of the two subdomains of the SARS-CoV 3C-like protease (which belongs to the serine protease superfamily). Nsp9 has presumably evolved from a protease. The crystal structure suggests that the protein is dimeric and this was confirmed by analytical ultracentrifugation and light scattering. Nsp9 binds RNA and appears to interact with nsp8. The SPINE structural and functional analyses indicate that nsp9 may play multiple roles in the replicative cycle of coronaviruses. Its interaction with other proteins may be essential for the formation of the viral replication complex together with its ability to interact with RNA (in the absence of other proteins). The loops presented by the β-barrel may principally confer the RNA-binding capacity *via* non-specific interactions, while the C-terminal β-hairpin and helix, which display a greater conservation across coronaviruses, are likely to be involved in dimerization and interaction with other proteins.

4.7. Vaccinia A41L

To date Oxford have solved four VACV protein structures, one of which, A41L, shares sequence similarity to several chemokine-binding proteins and is thought to play a role in reducing inflammatory responses to viral infection. Mutants of VACV deficient in A41L are readily cleared by the immune system and the A41L protein has also been shown to reduce the infiltration of inflammatory cells into infected areas of host immune systems in animal models. Oxford produced and purified A41L using the standard pipeline for expression in *E. coli*, followed by a generic refolding protocol, as mentioned above. Refolded protein was quality-assessed by MS. Crystal screens and optimizations were performed using the HTP protocols developed in Oxford (Walter *et al.*, 2003; Brown *et al.*, 2003; Walter *et al.*, 2005). The structure was solved by the MAD technique at the UK beamline BM14 (Grenoble) using SeMet-labelled A41L protein (M. Bahar, unpublished work). A41L is a single-domain protein with a core fold that adopts a distinct β-sandwich topology. The β-sandwich is defined by two β-sheets arranged parallel to each other (Fig. 3) and connected by an array of long loops. VACV A41L is structurally similar to the 35K chemokine-binding protein of the related cowpox virus. Work is under way to identify the chemokines to which the molecule binds, but it is clear from the structure that the interactions are somewhat unusual since the protein does not possess the overwhelmingly negatively charged surface usually associated with chemokine-binding proteins.

4.8. Epstein–Barr virus dUTPase

The EBV project in Grenoble attempted a structural genomics approach to enzymes of this virus; however, most aspects of the activity were not performed in HTP mode (Tarbouriech *et al.*, 2006). The exception to this was the use for crystal screening of a Cartesian robot dispensing nanolitre drops, which had a significant impact on the project (Berry *et al.*, 2006) and led to the determination of the structures of four proteins. Of these, EBV dUTPase is of particular interest. dUTPases are ubiquitous enzymes hydrolyzing dUTP in order to maintain a low intracellular concentration of dUTP and to minimize its incorporation in DNA. Monomeric dUTPases only occur in herpesviruses, whereas other organisms encode related trimeric or unrelated dimeric forms of this enzyme. For trimeric dUTPases, the three different subunits contribute five conserved motifs to each of the three active sites located at the subunit interfaces. The structure of EBV dUTPase represents the first example of a monomeric dUTPase (Tarbouriech *et al.*, 2005). The structure was determined in complex with the reaction product dUMP and the substrate analogue α,β-imino-dUTP. EBV dUTPase (256 residues) consists of two domains, each structurally similar to the subunit of the trimeric dUTPases, which contribute motif III, and motifs I, II and IV, respectively, to the creation of a single active site (Fig. 3). The C-terminal motif V is largely disordered in the solved structures but is tethered near the active site by an unexpected disulfide bridge between Cys4 and Cys246. The enzyme is a rare example of an evolution from a multimeric to a monomeric protein. Presumably, a single gene-duplication event led to a molecule which maintained an active site extremely similar in structure to those of the trimeric enzymes. It is to be hoped that such subtle differences between the active sites of the viral and the human enzymes can be exploited in the design of specific inhibitors.

5. Concluding remarks

The SPINE Project has contributed to a shift in the way structural biology is now being carried out in Europe through the democratization of the use of new technologies and the development of novel strategies at various steps of the structure-determination pipeline. Since SPINE is driven by the notion of selecting 'high-value human health targets', it was natural that a number of human pathogens were targeted for analysis, from bacteria and viruses that cause 'established' human diseases to newly emerging threats to human health (SARS-CoV). Overall, SPINE has solved a substantial number of structures from these targets (over 220, including 65 complexes), some of which are being evaluated as possible new drug targets. However, we have found no generic solution to the traditional problem of protein solubility; rather, the problem has been significantly ameliorated by the sheer numbers of potential constructs that can be screened through the establishment and development of parallel cloning and expression technologies both in bacterial and eukaryotic expression systems. For the future it seems that more comprehensive library-based approaches (sometimes termed 'directed evolution') may have a substantial impact in providing a much more extensive screening for soluble constructs and one such approach, the ESPRIT method (see Alzari *et al.*, 2006), is being applied to several viral targets of high value, including a SARS-CoV protein, in part as a collaboration between Oxford and Grenoble (C. Meier, unpublished results). There is considerable scope for further developments.

The analysis of a set of relatively straightforward bacterial targets was key to the benchmarking of the techniques used and developed in SPINE, especially for protein overexpression and purification. The York/Oxford collaboration on *B. anthracis* (Au *et al.*, 2006) identified a number of problems in the first generation of protocols, such as the insolubility of certain Gateway constructs. The failure to overexpress significant amounts of soluble protein could easily have been put down to the challenging nature of human and human viral targets if only these problems had been used to test pipelines. The work on bacterial targets pinpointed the problems and provided a convenient and reliable benchmark for their resolution. In addition, the *B. anthracis* project has led to the determination of 45 structures to date, which demonstrates the high success rate which can be achieved with the refined pipelines (up to 30%). In addition, the bacterial projects facilitated the ready sharing of knowledge, technologies and technique, which has contributed greatly to the

establishment of European HTP activities, and the dissemination of technologies to the wider community.

The production of soluble proteins for certain eukaryotic viral targets was poor in *E. coli* and in terms of amenability to overexpression in soluble form at suitable concentrations for NMR or protein crystallography these proteins behave more like human proteins. For these targets, eukaryotic expression systems such as baculovirus and mammalian cells provide powerful alternative vehicles for the production of soluble protein. It is interesting to note that the pattern of soluble protein expression across different viruses can vary markedly. Whereas herpes viral structural proteins were in general intractable in *E. coli*, there was a higher success rate with proteins from viruses such as VACV (from a small target set of ten, three crystal structures have been solved).

This work was funded by the European Commission as SPINE, Structural Proteomics In Europe, contract No. QLG2-CT-2002-00988 under the Integrated Programme 'Quality of Life and Management of Living Resources'.

References

AB, E. *et al.* (2006). *Acta Cryst.* D**62**, 1150–1161.
Albeck, S. *et al.* (2006). *Acta Cryst.* D**62**, 1184–1195.
Alcami, A., Khanna, A., Paul, N. L. & Smith, G. L. (1999). *J. Gen. Virol.* **80**, 949–959.
Alekshun, M. N. & Levy, S. B. (1997). *Antimicrob. Agents Chemother.* **41**, 2067–2075.
Alzari, P. *et al.* (2006). *Acta Cryst.* D**62**, 1103–1113.
Aricescu, A. R., Assenberg, R. *et al.* (2006). *Acta Cryst.* D**62**, 1114–1124.
Aricescu, A. R., Lu, W. & Jones, E. Y. (2006). *Acta Cryst.* D**62**, 1243–1250.
Au, K. *et al.* (2006). *Acta Cryst.* D**62**, 1267–1275.
Banci, L., Bertini, I., Ciofi-Baffoni, S., D'Onofrio, M., Gonnelli, L., Marhuenda-Egea, F. C. & Ruiz-Duenas, F. J. (2002). *J. Mol. Biol.* **317**, 415–429.
Banci, L., Bertini, I., Ciofi-Baffoni, S., Gonnelli, L. & Su, X. C. (2003*a*). *J. Mol. Biol.* **331**, 473–484.
Banci, L., Bertini, I., Ciofi-Baffoni, S., Gonnelli, L. & Su, X. C. (2003*b*). *J. Biol. Chem.* **278**, 50506–50513.
Banci, L. *et al.* (2006). *Acta Cryst.* D**62**, 1208–1217.
Banci, L. & Rosato, A. (2003). *Acc. Chem. Res.* **36**, 215–221.
Bellinzoni, M. & Riccardi, G. (2003). *Trends Microbiol.* **11**, 351–358.
Berrow, N. S. *et al.* (2006). *Acta Cryst.* D**62**, 1218–1226.
Berry, I. M., Dym, O., Esnouf, R. M., Harlos, K., Meged, R., Perrakis, A., Sussman, J. L., Walter, T. S., Wilson, J. & Messerschmidt, A. (2006). *Acta Cryst.* D**62**, 1137–1149.
Betton, J. M. (2004). *Biochimie*, **86**, 601–605.
Brown, J. *et al.* (2003). *J. Appl. Cryst.* **36**, 315–318.
Calvo, J. M. & Matthews, R. G. (1994). *Microbiol. Rev.* **58**, 466–490.
Daugelat, S., Kowall, J., Mattow, J., Bumann, D., Winter, R., Hurwitz, R. & Kaufmann, S. H. (2003). *Microbes Infect.* **5**, 1082–1095.
Donnelly, C. A., Ghani, A. C., Leung, G. M., Healey, A. J., Fraser, C., Riley, S., Abu-Raddad, L. J., Ho, L. M., Thatch, T. Q., Chau, P., Chau, K. P., Lam, T. H., Tse, L. Y., Tsang, T., Liu, S. H., Kong, J. H., Lau, E. M., Ferguson, N. M. & Anderson, R. M. (2003).*Lancet*, **361**, 1761–1766.
Egloff, M. P., Benarroch, D., Selisko, B., Romette, J. L. & Canard, B. (2002). *EMBO J.* **21**, 2757–2768.
Egloff, M. P., Ferron, F., Campanacci, V., Longhi, S., Rancurel, C., Dutartre, H., Snijder, E. J., Gorbalenya, A. E., Cambillau, C. & Canard, B. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 3792–3796.
Espinal, M. A. (2003). *Tuberculosis*, **83**, 44–51.
Geerlof, A. *et al.* (2006). *Acta Cryst.* D**62**, 1125–1136.
Gubser, C., Hue, S., Kellam, P. & Smith, G. L. (2004). *J. Gen. Virol.* **85**, 105–117.
Harris, N. V., Weiss, N. S. & Nolan, C. M. (1986). *Am. J. Public Health*, **76**, 407–411.
Kazwala, R. R., Collins, J. D., Hannan, J., Crinion, R. A. & O'Mahony, H. (1990). *Vet. Rec.* **126**, 305–306.
Kelly, D. J. (2001). *Symp. Ser. Soc. Appl. Microbiol.* **30**, 16S–24S.
Kobayashi, M., Kubota, M. & Matsuura, Y. (1999). *Acta Cryst.* D**55**, 931–933.
Kuiken, T., Fouchier, R., Rimmelzwaan, G. & Osterhaus, A. (2003). *Curr. Opin. Biotechnol.* **14**, 641–646.
Kurinov, I. V., Mao, C., Irvin, J. D. & Uckun, F. M. (2000). *Biochem. Biophys. Res. Commun.* **275**, 549–552.
Ma, Q., Zhao, X., Nasser Eddine, A., Geerlof, A., von Kries, J. P., Li, X., Cronan, J. E., Kaufmann, S. H. E. & Wilmanns, M. (2006). Submitted.
Marra, M. A. *et al.* (2003). *Science*, **300**, 1399–1404.
Miranda, A. B. de, Alvarez-Valin, F., Jabbari, K., Degrave, W. M. & Bernardi, G. (2000). *J. Mol. Evol.* **50**, 45–55.
Müller, A., Thomas, G. H., Horler, R., Brannigan, J. A., Blagova, E., Levdikov, V. M., Fogg, M. J., Wilson, K. S. & Wilkinson, A. J. (2005). *Mol. Microbiol.* **57**, 143–155.
Parkhill, J. *et al.* (2000). *Nature (London)*, **404**, 502–506.
Patterson, H. M., Brannigan, J. A., Cutting, S. M., Wilson, K. S., Wilkinson, A. J., AB, E., Diercks, T., Folkers, G., de Jong, R., Truffault V. & Kaptein. R. (2005). *J. Biol. Chem.* **280**, 36214–36220.
Peiris, J. S., Chu, C. M., Cheng, V. C., Chan, K. S., Hung, I. F., Poon, L. L., Law, K. I., Tang, B. S., Hon, T. Y., Chan, C. S. Chan, K. H., Ng, J. S., Zheng, B. J., Ng, W. L., Lai, R. W., Guan, Y. & Yuen, K.Y. (2003). *Lancet*, **361**, 1767–1772.
Rayment, I. (1997). *Methods Enzymol.* **276**, 171–183.
Ren, J., Sainsbury, S., Berrow, N., Alderton, D., Nettleship, J., Stammers, D. K., Saunders, N. J. & Owens, R. J. (2005). *BMC Struct. Biol.* **5**, 13–21.
Rota, P. A. *et al.* (2003). *Science*, **300**, 1394–1399.
Schnell, R., Sandalova, T., Hellman, U., Lindqvist, Y. & Schneider, G. (2005). *J. Biol. Chem.* **280**, 27319–27328.
Schubot, F. D. & Waugh, D. S. (2004). *Acta Cryst.* D**60**, 1981–1986.
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.
Smith, G. L., Symons, J. A. & Alcami, A. (1999). *Arch. Virol. Suppl.* **15**, 111–129.
Smith, G. L., Symons, J. A., Khanna, A., Vanderplasschen, A. & Alcami, A. (1997). *Immunol. Rev.* **159**, 137–154.
Smith, V. P. & Alcami, A. (2000). *J. Virol.* **74**, 8460–8471.
Snijder, E. J., Bredenbeek, P. J., Dobbe, J. C., Thiel, V., Ziebuhr, J., Poon, L. L., Guan, Y., Rozanov, M., Spaan, W. J. & Gorbalenya, A. E. (2003). *J. Mol. Biol.* **331**, 991–1004.
Spinelli, S., Desmyter, A., Verrips, C. T., de Haard, H. J. W., Moineau, S. & Cambillau, C. (2006). *Nature Struct. Mol. Biol.* **13**, 85–89.
Sutton, G., Fry, E., Carter, L., Sainsbury, S., Walter, T., Nettleship, J., Berrow, N., Owens, R., Gilbert, R., Davidson, A., Siddell, S., Poon, L. L., Diprose, J., Alderton, D., Walsh, M., Grimes, J. M. & Stuart, D. I. (2004). *Structure*, **12**, 341–353.
Svedhem, A. & Kaijser, B. (1981). *J. Infect.* **3**, 37–40.
Tarbouriech, N., Buisson, M., Géoui, T., Daenke, S., Cusack, S. & Burmeister, W. P. (2006). *Acta Cryst.* D**62**, 1276–1285.
Tarbouriech, N., Buisson, M., Seigneurin, J. M., Cusack, S. & Burmeister, W. P. (2005). *Structure*, **13**, 1299–1310.
Tettelin, H. *et al.* (2000). *Science*, **287**, 1809–1815.
Tortorella, D., Gewurz, B. E., Furman, M. H., Schust, D. J. & Ploegh, H. L. (2000). *Annu. Rev. Immunol.* **18**, 861–926.
Tremblay, D. M., Tegoni, M., Spinelli, S., Campanacci, V., Blangy, S., Huyghe, C., Desmyter, A., Labrie, S., Moineau, S. & Cambillau, C. (2006). *J. Bacteriol.* **188**, 2400–2410.

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. *et al.* (2005). *Acta Cryst.* D**61**, 651–657.
Walter, T. S., Meier, C., Assenberg, R., Au, K.-F., Ren, J., Verma, A., Nettleship, J. E., Owens, R. J., Stuart, D. I. & Grimes, J. M. (2006). In the press.
Wyszynska, A., Raczko, A., Lis, M. & Jagusztyn-Krynicka, E. K. (2004). *Vaccine*, **22**, 1379–1389.
Ziebuhr, J. & Siddell, S. (2002). *Encyclopedia of Life Sciences*, pp. 190–198. London: Stockton Press.
Zimhony, O., Vilcheze, C. & Jacobs, W. R. Jr (2004). *J. Bacteriol.* **186**, 4051–4055.