Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

SPINE bioinformatics and data-management aspects of high-throughput structural biology

S. Albeck,[a] P. Alzari,[b] C. Andreini,[c] L. Banci,[c] I. M. Berry,[d] I. Bertini,[c] C. Cambillau,[e] B. Canard,[e] L. Carter,[d] S. X. Cohen,[f] J. M. Diprose,[d] O. Dym,[a] R. M. Esnouf,[d] C. Felder,[a] F. Ferron,[e] F. Guillemot,[b] R. Hamer,[d] M. Ben Jelloul,[f] R. A. Laskowski,[g] T. Laurent,[g] S. Longhi,[e] R. Lopez,[g] C. Luchinat,[c] H. Malet,[e] T. Mochel,[h] R. J. Morris,[g] L. Moulinier,[h] T. Oinn,[g] A. Pajon,[g] Y. Peleg,[a] A. Perrakis,[f] O. Poch,[h] J. Prilusky,[a] A. Rachedi,[g] R. Ripp,[h] A. Rosato,[c] I. Silman,[a] D. I. Stuart,[d] J. L. Sussman,[a] J.-C. Thierry,[h] J. D. Thompson,[h] J. M. Thornton,[g] T. Unger,[a] B. Vaughan,[g] W. Vranken,[g] J. D. Watson,[g] G. Whamond[g] and K. Henrick[g]*

[a]The Israel Proteomics Center, The Department of Structural Biology, The Weizmann Institute of Science, Rehovot 76 100, Israel, [b]Unité de Biochimie Structurale, Institut Pasteur, 25–28 Rue du Dr Roux, 75724 Paris CEDEX 15, France, [c]CIRMMP, CERM, Via Sacconi 6, 50019 Sesto Fiorentino, Italy, [d]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, England, [e]Architecture et Fonction des Macromolécules Biologiques, UMR6098 CNRS/Université Aix-Marseille I and II, 31 Chemin Joseph Aiguier, 13402 Marseille CEDEX 20, France, [f]Division of Molecular Carcinogenesis, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands, [g]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England, and [h]Institut de Génétique et de Biologie Moléculaire et Cellulaire, 1 Rue Laurent Fries, BP 163, 67404 Illkirch CEDEX, France

Correspondence e-mail: henrick@ebi.ac.uk

Received 28 November 2005
Accepted 31 July 2006

SPINE (Structural Proteomics In Europe) was established in 2002 as an integrated research project to develop new methods and technologies for high-throughput structural biology. Development areas were broken down into workpackages and this article gives an overview of ongoing activity in the bioinformatics workpackage. Developments cover target selection, target registration, wet and dry laboratory data management and structure annotation as they pertain to high-throughput studies. Some individual projects and developments are discussed in detail, while those that are covered elsewhere in this issue are treated more briefly. In particular, this overview focuses on the infrastructure of the software that allows the experimentalist to move projects through different areas that are crucial to high-throughput studies, leading to the collation of large data sets which are managed and eventually archived and/or deposited.

1. Introduction

Bioinformatics as originally defined is the application of informatics techniques to biological systems and this general data-management and analysis problem is commonly regarded as being of central importance to large-scale genomic projects such as structural genomics (Burley *et al.*, 1999; Stevens *et al.*, 2001). Although not strictly a structural genomics project itself, the remit of Structural Proteomics In Europe (SPINE) in developing high-throughput methods for structural biology and the distributed nature of the project suggests that its data-management issues will be similarly critical. However, while it is generally accepted that the current principal bottlenecks in structural biology are the production of soluble protein and its crystallization, bioinformatics and data management are much more than a necessary housekeeping exercise. The experience of SPINE demonstrates that by defining a comprehensive data model for the process of structural biology and then implementing it in a form useful to non-bioinformaticists, although challenging, can yield extensive scientific added value. Recording both failure and success in the process allows objective analyses, including data mining, to be brought to bear on the bottlenecks in the overall process. The drive to develop an infrastructure to give structural biologists a simple yet comprehensive view that allows decisions to be reached over target selection and experimental strategy underlies the (bio)informatics work within SPINE. The issues addressed include the application of bioinformatics to target selection and analysis, the development of laboratory information-management systems (LIMS), the development of information-exchange systems between SPINE partners and the public presentation of the results of SPINE. This article gives an overview of these developments, particularly as they pertain to high-throughput initiatives. Fig. 1 shows an overview of the bioinformatics and data-management requirements for SPINE.

All structural genomics and high-throughput structural biology initiatives have included components aimed at data management: some try to reproduce particular facets of current laboratory and notebook practice, while others have grander visions of altering work practice wholesale. The development of novel bioinformatical analyses *per se* has not usually been a priority, but several informatics projects start with target selection and aim to span all aspects of laboratory work including final report writing, *e.g.* SESAME (Zolnai *et al.*, 2003), SPINE2 (Goh *et al.*, 2003) and SPEX (Raymond *et al.*, 2004). Of these, the SESAME project has perhaps had the most impact and is in use in more than one structural genomics consortium. Packages targeted at protein-production stages have included LISA (Haebel *et al.*, 2001), HalX (Prilusky *et al.*, 2005) and MOLE (Morris, Wood *et al.*, 2005), the latter two being from groups closely associated with SPINE. Downstream into crystallization and structure determination, CLIMS (Fulton *et al.*, 2004), PlateDB and the *Vault* software (Mayo *et al.*, 2005) and Xtrack (Harris & Jones, 2002; http://xray.bmc.uu.se/xtrack) offer tools for these localized data-management hotspots, respectively. Methods for deposition and presenting of target progress information have also benefited from the drive towards high throughput: these include the Protein Data Bank (PDB; Berman *et al.*, 2000) and the Macromolecular Structure Database (MSD; Boutselakis *et al.*, 2003) deposition tools and, in particular, the TargetDB database (Chen *et al.*, 2004; http://targetdb.pdb.org/) for publicly recording the activity of structural genomics projects.

One unique feature of SPINE has been the wide range of biomedically important protein targets ranging from viruses and bacteria to human protein targets of potential pharmaceutical interest. Working with such difficult target sets naturally impacted on the priorities attached to different types of informatics developments. In the area of target selection, effort concentrated on the development of sophisticated genome-sequence analysis and annotation tools, where selected protein targets or families were characterized and domain boundaries estimated. Several bioinformatics resources were developed specially to address problems posed by complex targets, such as VaZyMolO, a viral genome database developed by the Marseille partners, and there has been a special focus on native disorder-prediction systems such as *RONN* (Oxford/Exeter) and *FoldIndex* (Weizmann) as described in Esnouf *et al.* (2006). An important requirement for any target-annotation system is that information should be current (and updated regularly) in order to draw on the vast amounts of information being produced by ongoing projects such as complete genome sequencing, functional proteomics and 'interactomics' projects. Systems such as *PipeAlign* (Strasbourg), *SeqAlert* (Weizmann) and *OPTIC* (Oxford) were developed to address this aspect of the target-annotation process. Other developments were tailored to highly focused target sets such as those from mycobacteria (Pasteur) and metalloproteins (Florence).

Although widely considered essential for high-throughput and distributed projects, data management of experimental results using LIMS is at a very early stage of development. SPINE has played a central role in European efforts to provide an effective data-management solution covering experiments. It has supported the creation of a universal dictionary for describing experiments associated with protein production (the Protein Production Data Model curated by the EBI partner; Pajon *et al.*, 2005) and is working to extend the scope of this system to cover other aspects of structural proteomics. SPINE has also played a leading role in bringing together groups developing piecemeal solutions to data-management problems, brokering an agreement to develop a common framework for LIMS across Europe. These efforts are now beginning to bear fruit in the PIMS project (http://www.pims-lims.org) that is being developed by several SPINE partners in collaboration with others.

Following protein production, the processes leading to structure determination are increasingly stable and well defined, the data-analysis and management tools reflecting this with increasing sophistication. The rapid spread of small-volume crystallization trial robots combined with automated plate-storage and imaging systems has led to the development of high-thoughput, high-capacity, web-based management tools, such as the *Vault* software developed in Oxford (Mayo *et al.*, 2005), which is now being incorporated into PIMS developments as a generally applicable crystallization trial-management system (see Berry *et al.*, 2006). The Oxford infrastructure has also served to drive the development of automated crystal recognition software at York (Wilson, 2002, 2004; Berry *et al.*, 2006). Automation of the stages between crystallization in the home laboratory and synchrotron data collection have been addressed by a separate SPINE workpackage (WP6) and include significant data-management aspects which have been covered by links to automation projects such as eHTPX, ISPyB and DNA (Beteva *et al.*, 2006). Combined with technological developments (*e.g.* automated sample changers) and standardization (*e.g.* in pin design and barcodes), these initiatives are now beginning to deliver better science and the real prospect of standardized remote data collection at multiple synchrotron sites (Beteva *et al.*, 2006; Cipriani *et al.*, 2006). Finally, automation of structure solution was covered by SPINE workpackage 7, in which the emphasis has been to bring developers together and to provide test data rather than developing novel methods or attempting to manage the data during the structure-determination process (Bahar *et al.*, 2006). However, the collaborative frameworks fostered by SPINE offer the opportunity to address extending the data-management framework into structure solution by bringing together synchrotrons, methods developers and laboratories with high-throughput ambitions.

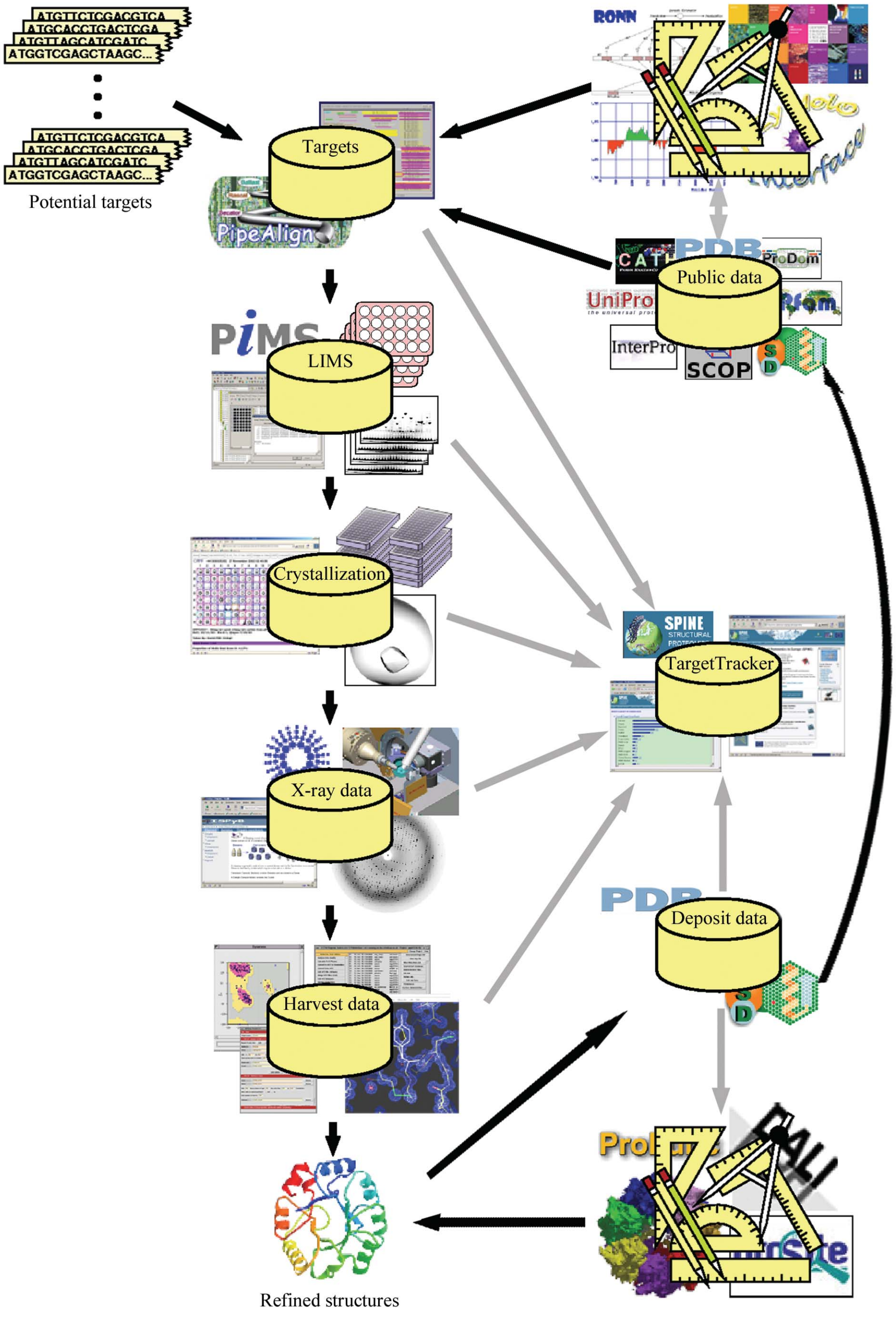


Figure 1
A schematic diagram showing the SPINE high-throughput structure-determination pipeline, highlighting the role played by bioinformatics and data management. The yellow drums indicate databases and the primary flow of information is indicated by black arrows. Potential target sequences are annotated using bioinformatics tools and collated into target databases. Selection of targets for expression and construct design is guided by bioinformatics data. Data on the protein-production process are recorded by a LIMS and other informatics tools. For crystallography, purified soluble protein samples enter crystallization trials using robotic setup, storage and imaging procedures, all of which are managed by custom databases and software interfaces. Data are collected from crystals either in-house or at synchrotron sources requiring crystal-shipping management and automated beamline operation and record keeping. Phasing and structure refinement are also increasingly automated. Refined structures are annotated by reference to the literature and using automated tools. These structures are then deposited in the public structure database, which contributes towards improving sequence and structure-annotation tools, feeding back into more informed target selection.

2. Software components

2.1. The SPINE website

The SPINE website, http://www.spineurope.org/, aims to provide an information portal for all activities relating to the SPINE project and addresses three key audiences. Firstly, it provides information to the general public about all public aspects of the project including its background, partner contact details, project work-packages, an up-to-date scoreboard and list of structures, key publications, forthcoming meetings, data and tools generated by the project, project-derived protocols and news. Secondly, it provides a repository of information for members of SPINE to store and share private information such as meeting presentations, private protocols, unpublished structures and administrative information. Thirdly, it provides extended controlled access to members of the SPINE industrial platform. The website is built around the free open-source bulletin board system PHPBB (http://www.phpbb.com/) and it uses the common internet portal framework provided by PHP (Zandstra, 2004; http://www.php.net) and MySQL (DuBois, 1999), upon which a custom-built content-management system was built.

2.2. The SPINE targets database

SPINE targets are held in a database at the EBI and are available in a format defined by an XML document-type definition file (`target.dtd`; http://www.ebi.ac.uk/msd-srv/msdtarget/). The database is populated with data provided by SPINE partners either using a Java loader (the EBI target manager, described below) or online through a target register (http://www.ebi.ac.uk/msd-srv/msdtarget/cgi-bin/spnreg/login.pl) which can be run in-house for ORACLE or PostgreSQL (Matthew & Stones, 2001) databases. SPINE targets are annotated with data-integration tools that access public databases such as Medline (Macleod, 2002), EMBL (Hamm & Cameron, 1986), GO (The Gene Ontology Consortium, 2001) and PFAM (Bateman *et al.*, 2004). Information is retrieved through web-based tools for searching and tracking, including SpineSearch (http://www.ebi.ac.uk/msd-srv/msdtarget/spine/SPINEindexp.html), SpineStatus (http://www.ebi.ac.uk/msd-srv/msdtarget/status.html), SpineScoreboard (http://www.spineurope.org/page.php?page=scoreboard), SpineStructuralGalleries (http://www.spineurope.org/page.php?page=structures) and SpineAlert (http://www.ebi.ac.uk/msd-srv/msdtarget/MSDtargetsAlert/).

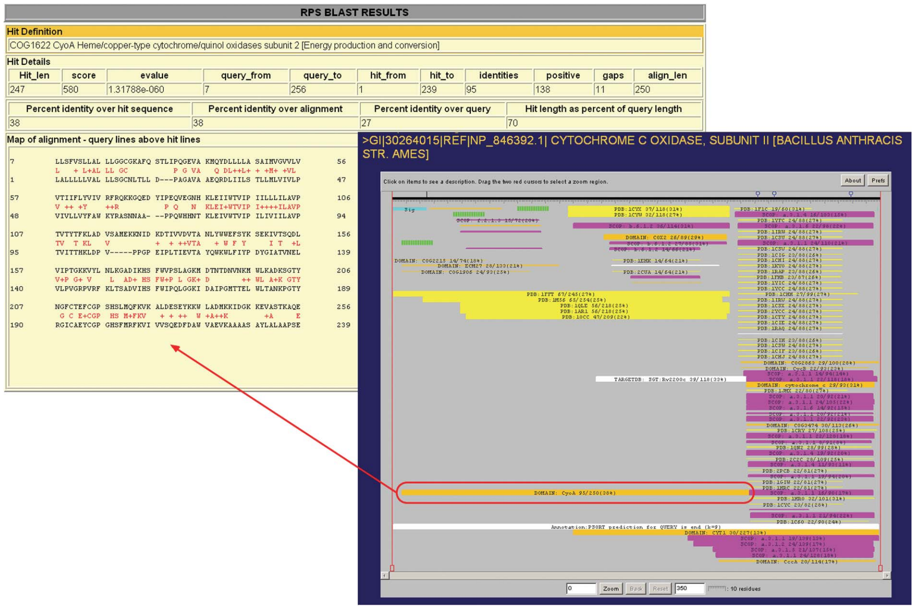
A customizable target scoreboard service (http://www.ebi.ac.uk/msd-srv/msdtarget/cgi-bin/spnreg/login.pl) is provided to enable multiple plots of target statistics against target status and species data. The EBI target-tracking system offers an efficient means of tracking the status of targets that are products of a single gene, following the lead of TargetDB (Chen *et al.*, 2004). However, this framework cannot handle protein complexes, which form a significant proportion of SPINE targets. To complement the SPINE schema (http://www.ebi.ac.uk/msd-srv/msdtarget/target.dtd), an extended schema (http://www.ebi.ac.uk/msd-srv/msdtarget/spine/new_SPINE_schema/) was developed to deal with not only simple targets but also complexes and targets that are either cocrystallized or co-expressed and allows referencing between targets. The Amsterdam partners have designed a database and data-capture system that allows individual targets to be combined as complexes. A more extensive overview of that work is reported by Romier *et al.* (2006).

The EBI target-manager application was developed using WebObjects, a system that uses Java to process queries and build webpages. The application was designed to allow for easy data entry by small laboratories. The system allows laboratory personnel to store and update data relevant to their proteins and the status of their projects at any time during their work, including dates, access privileges and contact information. The protein data are submitted *via* standard XML forms from the UniProt database (Bairoch *et al.*, 2005). All data are initially stored in a local database and a scheduled weekly job builds an XML file in the EBI SPINE standard, which is automatically uploaded by the EBI to keep the status of targets up to date.

The targets database currently tracks information on 2260 SPINE targets distributed 66 and 34% between workpackages on pathogens and human proteins. A detailed breakdown of progress in these areas is reported in Alzari *et al.* (2006) and Banci *et al.* (2006). Of the 20 SPINE partners contributing targets, some have automated or semi-automated procedures, but the majority load targets manually with the EBI target manager.

2.3. Target selection and construct design

This activity divides into two distinct phases. Firstly, in order to understand the potential biomedical role of a candidate protein and hence to decide if the candidate actually warrants study, diverse data including the source organism, predicted domain organization, data on splicing variants, structures of related sequences and the significance of any known mutations must be organized into an information network and presented to the experimentalist. Secondly, for each particular protein (or complex) that appears interesting it is then important to determine whether one or more constructs can be defined that are likely to prove amenable to structural analysis (*e.g.* do they contain likely transmembrane helices or significant regions of disorder?). Within SPINE, several partners have built on their experience in these areas to provide tools addressing both of these phases. Although there is some commonality between the requirements of different laboratories and there was open exchange of ideas, no attempt was made to construct a single platform incorporating all these requirements. Nevertheless, the powerful tools described below are publicly available and the SPINE-adopted protein-production data model (Pajon *et al.*, 2005) may provide a framework for assembling the next generation of more unified software.



RPS BLAST RESULTS

Hit Definition

COG1622 CyoA Heme/copper-type cytochrome/quinol oxidases subunit 2 [Energy production and conversion]

Hit Details

Hit_len	score	evalue	query_from	query_to	hit_from	hit_to	identities	positive	gaps	align_len
247	580	1.31788e-060	7	256	1	239	95	138	11	250

Percent identity over hit sequence	Percent identity over alignment	Percent identity over query	Hit length as percent of query length
38	38	27	70

Map of alignment - query lines above hit lines

```
7          LLSFVSLLAL LLGGCGKAFQ STLIPQGEVA KMQYDLLLLA SAIMVGVVLV        56
           L   + L+AL LL GC          P G VA   Q DL++L+ + +M+ +VL
1          LALLLLLVAL LLSGCNLTLL D---PAGAVA AEQRDLIILS TLLMLVIVLP        47

57         VTIIFLYVIV RFRQKKGQED YIPEQVEGNH KLEIIWTVIP IILLLILAVP        106
           V ++ +Y    ++R          P Q   N  KLEI+WTVIP I++++ILAVP
48         VIVLLVYFAW KYRASNNAA- --PPQWHHNT KLEIVWTVIP ILIVIILAVP        94

107        TVTYTFKLAD VSAMEKKNID KDTIVVDVTA NLYWWEFSYK SEKIVTSQDL        156
           TV  T KL   V          +  + ++VTA  + W F Y     I T  +L
95         TVITTHKLDP V-----PPGP EIPLTIEVTA YQWKWLFIYP DYGIATVNEL        139

157        VIPTGKKVYL NLKGADIKHS FWVPSLAGKM DTNTDNVNKM WLKADKSGTY        206
           V+P G+ V    L  AD+ HS FW+P L GK+ D        ++ WL A+K GTY
140        VLPVGRPVRF KLTSADVIHS FWIPQLGGKI DAIPGMTTEL WLTANKPGTY        189

207        NGFCTEFCGP SHSLMQFKVK ALDESEYKKW LADMKKIDGK KEVASTKAQE        256
            G C E+CGP  HS M+FKV   + + ++  W +A++K          +A     E
190        RGICAEYCGP GHSFMRFKVI VVSQEDFDAW VAEVKAAAAS AYLALAAPSE        239
```

>GI|30264015|REF|NP_846392.1| CYTOCHROME C OXIDASE, SUBUNIT II [BACILLUS ANTHRACIS STR. AMES]

Click on items to see a description. Drag the two red cursors to select a zoom region.

Figure 2
Presentation of automated sequence analysis performed using the integrated bioinformatics suite developed at Oxford (OPAL annotation tool and OPTIC database). The schematic shows the *SEView*-based display for one of the proteins annotated as part of the joint Oxford/York study of proteins from *B. anthracis*. The target protein sequence runs across the top of the diagram and coloured bars indicate regions of the target sequence that show significant sequence similarity to sequences in bioinformatics/structural databases. The width of each bar is related to the degree of sequence similarity. The different colours indicate different sources of information, *e.g.* yellow, PDB entries; orange, domain definitions; magenta, SCOP definitions; white, TargetDB entries. Clicking on any bar will bring up further information. The example indicated by the red arrow shows the detailed comparison of the target sequence to a COG family.

2.3.1. Strasbourg. The Strasbourg partners have provided a web server for protein-family analysis (PipeAlign) incorporating target curation and validation protocols as well as automatic structure-based hierarchical multiple alignment analysis (Plewniak *et al.*, 2003). The platform is generally available (http://igbmc.u-strasbg.fr/PipeAlign/) and integrates a cascade of programs for the automatic collection of sequence information and the construction and validation of multiple alignments of protein families. A number of new analysis programs have been developed and integrated into the platform. *NorMD* (Thompson *et al.*, 2001) implements an objective function to measure the quality of a multiple alignment, *RASCAL* (Thompson *et al.*, 2003) is used for the correction and refinement of an alignment and *LEON* (Thompson *et al.*, 2004) evaluates the extent of homology between the sequences in the alignment. This cascade of programs has been integrated in the Strasbourg Gscope bioinformatics platform that allows automatic high-throughput data collection, cross-validation and analysis of heterogeneous information in a single integrated environment. The integration of the protein information in the context of the complete family provides the basis for the definition of the hierarchical relationships within and between subfamilies and for the reliable integration of all the structural and functional information available for the protein family. A new program, *MACSIM* (manuscript in preparation), has been developed whose primary goal is to validate the quality of the data recovered from public databases and to propagate this information to the target of interest. A separate web server (http://igbmc.u-strasbg.fr/vALId/) has been developed for the validation and curation of protein sequences, including the detection of genome-annotation errors and splicing errors (Bianchetti *et al.*, 2005). The Gscope platform has been used to perform PipeAlign and MACSIM analyses of all targets in the SPINE target database and an 'identity card' has been created for each potential target. These 'identity cards' are available in a standard XML data-exchange format *via* the SPINE web site. To save time and avoid mistakes, Strasbourg have completed this part of the pipeline by developing a combinatorial interface for primer design (R. Ripp, manuscript in preparation).

2.3.2. Oxford. The Oxford node has developed a single resource for protein and DNA analysis, the Oxford Protein Analysis Linker (OPAL), under which sits the Oxford Protein Target Information Collection (OPTIC), a database for storing the results of these analyses. OPAL incorporates both publicly available analysis tools and bespoke tools developed in-house. These tools include the calculation of various parameters from the sequence such as pI, GRAVY (**gr**and **av**erage of hydropath**y**; Kyte & Doolittle, 1982), extinction coefficient and rare-codon usage, the prediction of signal sequence and cellular location predictions, potential glycosylation sites, secondary-structure predictions, transmembrane region predictions and potential protease-cleavage sites. They also include *BLAST*-based (Altschul *et al.*, 1990) sequence comparisons against publicly available databases, including CDD (Marchler-Bauer *et al.*, 2005), PDB (Berman *et al.*, 2000), SCOP (Andreeva *et al.*, 2004) and SWISS-PROT (Boeckmann *et al.*, 2003). The interface to OPAL is through a publicly available web form (http://www.oppf.ox.ac.uk/opal/OPAL.php). One or more sequences can be entered in FASTA format or specified by GenBank accession No. (Benson *et al.*, 2005) and the choice of analyses is made through a set of tick boxes. The results are returned as a web page that users can save locally. Where hits are found in external databases, the page includes links to the relevant web site. The results can also be presented graphically using *SEView*, a Java applet for browsing molecular-sequence data (Junier & Bucher, 1998; see Fig. 2). For authorized users, there is an alternative input form which uses locally installed (and licenced) versions of the tools and local databases to create and store a full annotation in the OPTIC database. Not only is this process considerably faster, the stored annotations are also checked regularly and can generate user alerts (by e-mail). Visualization of OPTIC entries are based on the *SEView* applet. For the whole-genome scale the process of annotation can be trivially scripted.

The final phase of target selection is construct design, which has been largely automated in the OPTIC-linked *OPINE* application and is based on the Primer3 tool (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi) customized into a Windows DLL. *OPINE* provides the link, at two levels, between the bioinformatics-based exercise of target selection and experimental work. Firstly, based on its knowledge of which expression strategies are available in the laboratory, *OPINE* provides a simple-to-use tool for the design of primers. Secondly, at the data-management level it transfers data on the designed constructs into the LIMS to make it available for authorization (where the primer orders are created in appropriate 96-well format) and experimental work. The construct-design strategy has proven its reliability in the work on *Bacillus anthracis* (see Au *et al.*, 2006), with only two PCR failures for 114 Gateway (Invitrogen) pDEST14 and In-Fusion (Clontech) constructs, even when cloning out of long (>5 Mbp) genomic DNA. These construct-design tools are routinely used to define a set of constructs of high-value targets (as exemplified by the study of MICAL; Siebold *et al.*, 2005). The possibility of using data analysis to inform construct design is still at a very early stage. Within SPINE, this activity is exemplified by the Oxford/York analysis of the *B. anthracis* targets (Au *et al.*, 2006), which reinforces earlier findings from the Joint Center for Structural Genomics (JCSG; Page *et al.*, 2003; Stevens, 2004).

2.3.3. Weizmann Institute. The Weizmann node has developed, tested and opened for public use several web-based tools for automatic annotation.

(i) *FoldIndex* (http://bioportal.weizmann.ac.il/fldbin/findex) is a dynamic and interactive process that estimates the local and general probabilities that any query sequence will fold under specified conditions (Prilusky *et al.*, 2005; also considered further in Esnouf *et al.*, 2006).

(ii) *SeqAlert* (http://bioportal.weizmann.ac.il/salertb/main) is a sequence-alerting service that will periodically compare submitted sequences against sequences for known structures deposited with the PDB (Berman *et al.*, 2000) and targets recorded in TargetDB (Chen *et al.*, 2004). It also reports the

PubMed IDs of papers on proteins published over the previous 60 d that are related to the target sequence. One useful feature of *SeqAlert* is that it allows both bulk and e-mail-based requests.

(iii) *SeqFacts* (http://bip.weizmann.ac.il/sqfbin/seqfacts) is a tool for sequence identification, analysis, characterization and annotation.

A description of the interface between these tools and a wet laboratory LIMS is given in §2.4. As with other partners, there is the possibility of performing primer design as part of the LIMS, in this case using the bestPrimer server.

2.3.4. Florence. The Florence partners have a particular focus on metalloproteins and have created the CIRMMP data-storage system for methods of metalloprotein identification and their structural modelling in a genome-wide context (AB *et al.*, 2006). A high-throughput ligand-docking approach based on *AutoDock* (Morris *et al.*, 1998) has been implemented through this system. The question of whether a protein needs a metal ion for its function is a major challenge in proteomics (Bertini & Rosato, 2003). Expression and purification of a protein cannot always provide an answer since a metalloprotein may be isolated in the demetallated form or, conversely, a non-metalloprotein may be obtained associated with a spurious metal ion. In some cases, metal-binding capabilities can be inferred for an uncharacterized protein through homology to a known metalloprotein. Florence has developed a methodology for the identification of metalloproteins in genome data banks (Andreini *et al.*, 2004) through known metal-binding patterns (MBPs). A similar approach was implemented for the Metalloprotein Database, a collection of MBPs automatically extracted from the PDB (Castagnetto *et al.*, 2002). For each metalloprotein in the PDB an MBP is attached. The primary structure of the metalloprotein (the query) from the PDB and the corresponding MBP are used as input for a variant of *BLAST*, *PHI-BLAST* (Zhang *et al.*, 1998) to scan gene banks (or a complete genome sequence), where the ratio between the number of amino acids aligned and the length of the query sequence can be taken as a good indicator of false positives. The method has been applied to four different genomes (*Pyrococcus furiosus*, *Escherichia coli*, *Drosophila melanogaster* and *Homo sapiens*) and a number of putative copper-binding proteins have been identified on this basis. Metalloproteins can only bind metals made available by the tight control system of the cell, which thus selects the ion(s) actually bound (Banci & Rosato, 2003) and this must be borne in mind when assessing the bioinformatics-based analysis. This approach was integrated with libraries of metal-binding protein domains based on multiple sequence alignments of known metalloproteins (*e.g.* taken from Pfam; Bateman *et al.*, 2004) and by scanning the annotations of gene sequences. On this basis it was estimated that zinc-binding proteins constitute around 10% of the human proteome, comprising mainly transcription factors (Bertini *et al.*, unpublished data). The identification of metalloprotein families using these techniques expanded the choice of targets available for input to the integrated crystallography/NMR structure-determination pipeline in SPINE (Eiso *et al.*, 2006).

2.3.5. Marseille. The Marseille partners have developed an in-house database, VaZyMolO db (Ferron *et al.*, 2005), which organizes open reading frames (ORFs) from complete genomic sequences for negative single-stranded RNA viruses and for *Flaviviridae*, *Narnaviridae*, *Coronaviridae* and *Arteriviridae*. Data are taken from three public databases: the viral genomes resource from the NCBI (Wheeler *et al.*, 2001; http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/viruses.html), ICTVdb, which is a taxonomic db (http://ictvdb.mirror.ac.cn/ICTVdB/index.htm), and VIDA (Mar-Albà *et al.*, 2001; http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html). The ORFs are organized into homologous protein families, which are identified on the basis of sequence similarity. Conserved sequence regions of potential functional importance are identified and can be retrieved as sequence alignments. Taxonomic and functional classifications are used for all the proteins and protein families in the database. The proteins are preclassified on the basis of function (*i.e.* surface proteins, matrix proteins, non-structural proteins). When available, protein structures that are related to the families are also included. For each entry in the database additional information (bibliography, activity, mutation effect, protein interactions and cellular localization) can be retrieved and a map of disordered and transmembrane regions is generated. These latter characteristics are problematic in terms of protein expression, solubility and crystallizability and may need to be excluded in order to increase the chance of success in structure determination. In summary, the purpose of VaZyMolO is to define protein domains (rather than entire gene products) suitable for structural work. The kernel of VaZyMolO is CAZy (Coutinho & Henrissat, 1999; http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html). The integrated tools used for the analysis are *BLAST* (Altschul *et al.*, 1990), *PSI-BLAST* (Zhang *et al.*, 1998), *TM-HMM* (Sonnhammer *et al.*, 1998), *PONDR* (Romero *et al.*, 2001), *ProtParam* (Gasteiger *et al.*, 2005), hydrophobic cluster analysis (*HCA*; Callebaut *et al.*, 1997), threading programs exploiting protein homology (Kelley *et al.*, 2000), *PP* (Rost, 1996), *DPANN* (Reinhardt & Eisenberg, 2004), fold recognition (Alexandrov *et al.*, 1995), Hidden Markov models (Krogh *et al.*, 1994) and profile–profile sequence alignments (*FFAS*03; Jaroszewski *et al.*, 2005). Data-collection procedures are accessible *via* http://afmb.cnrs-mrs.fr/stgen/vazymolo.html.

VaZyMolO was developed to define viral protein modules that might be expressed in soluble and functionally active forms, thereby identifying candidate proteins for crystallization studies. More than 170 complete viral genome sequences (from negative- and positive-sense single-stranded RNA viruses) have been annotated into modules, which were first identified by homology search and then validated by the convergence of results from sequence-composition analyses, motif search, transmembrane-region search and domain definition. This approach was particularly important in the structural characterization of the $2'$-*O*-methyltransferase domain of Dengue virus (Egloff *et al.*, 2002). Similarly, the VaZyMolO-based approach was used to define 57 modules within the SARS genome, which were submitted to the high-throughput platform for expression, purification and crystallization in Marseille. This approach led to the rapid crystal structure determination of the non-structural protein Nsp9 of SARS virus (Egloff *et al.*, 2004; see also Alzari *et al.*, 2006).

2.3.6. Pasteur. Bioinformatical analysis has shown that functional information is lacking for about 40% of the proteins predicted in *Mycobacterium tuberculosis* and suggests that many of these proteins may participate in novel metabolic pathways or confer unique properties such as the ability to persist indefinitely in infected tissue. A comparative analysis of the deduced proteomes from *M. tuberculosis* and *M. leprae* was performed and led to the identification of a set of over 300 proteins that are exclusively found either in mycobacteria or in actinomycetes, but which have no counterparts in other organisms. Since their genes are conserved in the degraded genome of *M. leprae*, these proteins might be presumed to be important for survival and therefore could provide potential drug targets. Those gene products restricted to mycobacteria (including 267 genes whose protein products were predicted to have soluble domains) together with other proteins of known function that could represent potential drug targets constituted the initial target set selected by the Pasteur for the SPINE project. The list and information on progress is available from http://feu.sis.pasteur.fr/cgi-bin/WebObjects/MINISGP as well as from http://www.spineurope.org/targetlist/.

2.3.7. EBI. The EBI node has developed a prototype computational resource to facilitate the automatic selection of potential target constructs for protein structure determination (Whamond, unpublished results). Starting from a protein sequence or a UniProt accession code (with an optional sequence range), a potential construct is chosen without the need for extensive human interaction with the software. The process checks for existing structures related to the sequence and highlights regions of interest such as potential structural domains. In general, determination of an appropriate homologue to a given sequence for structure determination by molecular replacement (see Bahar *et al.*, 2006) has involved running a series of computer programs, with the output from each stage in the process being manually copied and pasted into the input for the next stage. Here, the processes are linked using common data formats such as the XML schemas defined by the DAS (http://biodas.org/) and efamily (http://www.efamily.org.uk/) projects. In order to implement the workflow, the Taverna workbench (Oinn *et al.*, 2004; http://taverna.sourceforge.net/) has been used to formulate the workflow and implement each of the processes as distributed web services. For example, while the secondary-structure prediction steps of the workflow may be carried out on servers at the EBI, sequence-domain retrieval from Pfam can be carried out simultaneously on a server at the Sanger Centre. The major advantage of this kind of approach is the possibility of shifting the processing load from a user's local machine to remote servers, while allowing simultaneous processing to occur. Another advantage of the Taverna system is that it is platform-independent. Users can edit the workflow to their specification and add new processes (such as in-house tools) with relative ease. In collaboration with the Midwest Consortium for Structural Genomics (MCSG), the EBI are also working with sophisticated pattern-recognition and classification techniques to predict the solubility of a protein from its sequence, an important additional aid in target selection and prioritization.

2.4. Wet-laboratory LIMS

High-throughput structural projects have the potential to create a vast amount of diverse data. The number and variety of the target proteins and macromolecular assemblies, the diversity of experimental processes and the abundance of new laboratory practices and protocols make it very hard to follow and analyze experiments without the use of informatics. The stages involved in the production of proteins for structural studies include selection and design of targets, PCR, cloning, recombinant expression (both small scale and scale-up, both incubation and fermentation and using a variety of expression

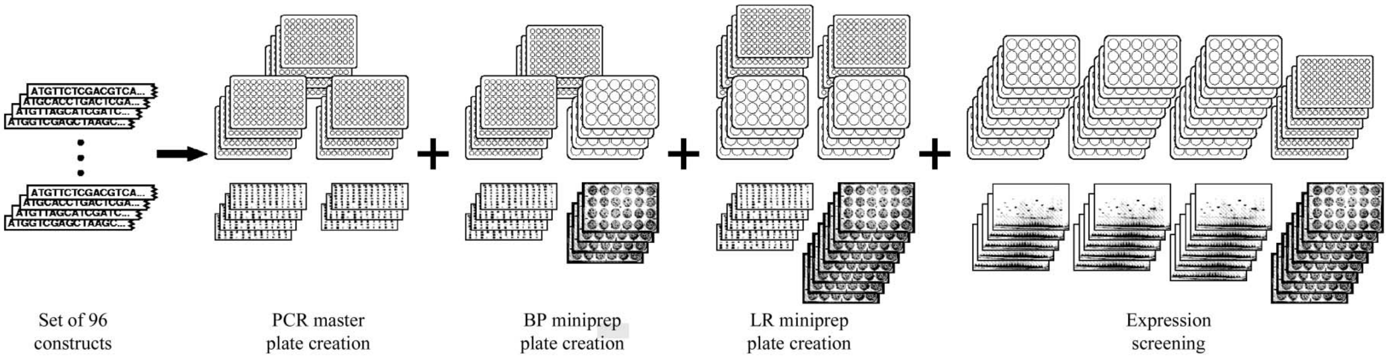
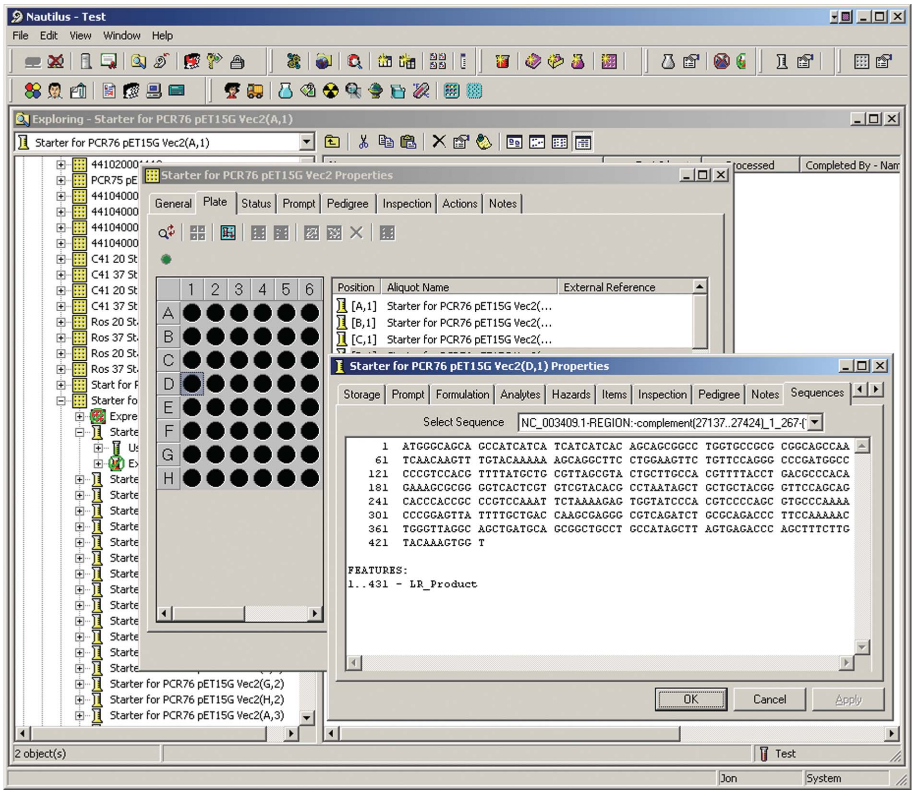


Figure 3
An indication of the scale of results generated by high-thoughput studies. The example considers the data generated from a single plate containing DNA for studying 96 constructs through PCR, the Gateway cloning protocol (Invitrogen) and small-scale expression screens using two cell lines with each of two experimental protocols. The top half of the figure shows the usage (and therefore requirement to record in a LIMS) of 96-well and 24-well plates, the bottom half shows the number of 96-lane agarose gels, 24-well colony-plate images and 26-lane SDS–PAGE gels that also need recording. In total, a set of 96 constructs uses 34 96-well plates and 36 24-well plates and generates 480 images of colony wells, 1536 lanes on agarose gels and 416 lanes on SDS–PAGE gels.

protocols) to produce native and labelled samples, purification and analytical characterization/quality assurance (QA). Furthermore, with high-value targets it is common to follow a strategy of multiple constructs and expression protocols. This approach generates a cascade of experimental results (Fig. 3) for which automated tracking becomes effectively essential. Informatics solutions to laboratory data-management problems are collectively referred to as laboratory information-management systems (LIMS). Commercial LIMS can be expensive and are usually highly tuned to specific (and unchanging) industrial processes, typically analytical and QA processes. This specificity makes them particularly poorly adapted to the fluid processes that characterize academic research. Thus, there is a clear need for an easy-to-use LIMS covering the processes of structural proteomics that is well adapted to use in the rapidly changing world of academic research.

In the absence of useful academic software and without in-house experience, the Oxford partners decided in 2001, prior to the start of SPINE, to use as a first implementation an adaptation of a commercially available LIMS, Nautilus (Thermo Electron Corporation). This has now been developed as far as practical linking back to target selection and primer design and covering plate layout, PCR, cloning, small-scale expression trials and crystallization trials. The middle stages, expression scale-up and purification, have proved very difficult to integrate for technical and design-based reasons and these are currently covered by fixed-format worksheets. The software is in everyday use and automated routes exist for the reporting of progress to the OPPF webpages and the SPINE web site. Nautilus is a two/three-tier LIMS underpinned by the Oracle RDBMS. Interaction with the database is *via* 'thick' clients (most calculations are performed locally on each client, with only database access to the server). The basic Nautilus schema is fixed and is built around a simple hierarchy of Sample-Group > Sample > Aliquot > Test > Result. Flexibility is introduced by allowing Aliquots to be split, pooled and linked recursively to other Aliquots in parent–child relationships. In the context of work in Oxford, any physically separate amount of any reagent is viewed as an Aliquot. Tests correspond to experiments, either characterizing an Aliquot or (as an Oxford-specific extension outside the original scope of Nautilus as an analytic LIMS) transforming one Aliquot into another, thereby allowing complex workflows to be modelled. Other useful features of Nautilus include a mechanism to represent experiments on plates of Aliquots, a patented workflow technology, a method for parsing information from instrument log files, a background processor for automatic execution of tasks, a way of incorporating custom-written code and a complex mechanism for maintaining database integrity and security. An example of the Nautilus user interface is shown in Fig. 4. When a construct design is authorized for laboratory work, its details are passed on to Nautilus. Through Nautilus, the construct designs are laid out in micro-titre plates, instructions to laboratory staff for the creation of PCR template plates are produced and correctly formatted orders for forward and reverse primer synthesis are generated. The PCR stage is fully integrated with the LIMS, with output from Mastercycler instruments (Eppendorf) being parsed automatically. The molecular-biology stages relevant to the Gateway system (BP and LR reactions; Invitrogen) as well as the steps involved in small-scale (*E. coli*) expression trials have also been modelled. Gels are a crucial part of the data accumulation in these stages and it is possible to incorporate data from scanning 96-well format gels into the database and to associate these scans directly with experiments (although this is not fully automated). As noted above,



Starter for PCR76 pET15G Vec2(A,1)

Position	Aliquot Name	External Reference
[A,1]	Starter for PCR76 pET15G Vec2(…	
[B,1]	Starter for PCR76 pET15G Vec2(…	
[C,1]	Starter for PCR76 pET15G Vec2(…	

Select Sequence: NC_003409.1-REGION:-complement(27137..27424)_1_267-(

```
    1  ATGGGCAGCA GCCATCATCA TCATCATCAC AGCAGCGGCC TGGTGCCGCG CGGCAGCCAA
   61  TCAACAAGTT TGTACAAAAA AGCAGGCTTC CTGGAAGTTC TGTTCCAGGG CCCGATGGCC
  121  CCCGTCCACG TTTTATGCTG CGTTAGCGTA CTGCTTGCCA CGTTTTACCT GACGCCCACA
  181  GAAAGCGCGG GGTCACTCGT GTCGTACACG CCTAATAGCT GCTGCTACGG GTTCCAGCAG
  241  CACCCACCGC CCGTCCAAAT TCTAAAAGAG TGGTATCCCA CGTCCCCAGC GTGCCCAAAA
  301  CCCGGAGTTA TTTTGCTGAC CAAGCGAGGG CGTCAGATCT GCGCAGACCC TTCCAAAAAC
  361  TGGGTTAGGC AGCTGATGCA GCGGCTGCCT GCCATAGCTT AGTGAGACCC AGCTTTCTTG
  421  TACAAAGTGG T

FEATURES:
1..431 - LR_Product
```

Figure 4
An example of the Oxford interface to the Nautilus LIMS (Thermo Electron Corporation). The back window shows a tree-based view onto the LIMS database showing all the plates known to the system. The middle window shows (some of) the properties for a particular plate (following user-interface standards, this view is obtained by right-clicking over the plate name and then selecting 'Properties') describing graphically the contents of each well in the plate. The front window displays (some of) the properties for an individual well (the window is again obtained with the expected mouse clicks), in this case the Oxford-developed tab displaying the sequence of the DNA sample in that particular well.

tracking of expression scale-up and purification has proved impractical to implement and development effort has now shifted to building this into a new framework (PIMS; see below).

The Weizmann node has developed and implemented a LIMS to cover its current needs. The system consists of a set of tools for automatic annotation of large-scale data set covering all aspects of going from gene to three-dimensional protein structure. This LIMS is intended as a 'laboratory notebook' and for tracking and evaluation of different methods. In close collaboration with Anne Poupon (University Paris-Sud, Orsay), the HalX LIMS has been extended (Prilusky *et al.*, 2005). For example, HalX is now capable of querying and retrieving information from remote servers. The first implementation of this web services feature was for the design of primers and is carried out routinely at the Weizmann by the bestPrimers server. In addition, Xtrack (Harris & Jones, 2002) is used to track crystallization and data-reduction steps. A custom-developed Perl application programming interface (API) for the protein-production data model (Pajon *et al.*, 2005) provides an open path to integrate with the wide selection of Perl-based bioinformatics tools and libraries that are available worldwide.

To develop LIMS further, three SPINE partners (Oxford, York and the EBI) have agreed to joint development of a free-to-use academic LIMS specially suited to protein production (the Protein Information Management System; PIMS; http://www.pims-lims.org). This development is starting from the protein-production data model, itself generated with input from several SPINE partners. The data model is emerging as a standardized method for experimental information interchange in the structural biology field and it forms the basis of HalX developments. Also building on this model, and in collaboration with several projects in this area (especially the UK BBSRC-funded eHTPX project), a web service portal (at Daresbury Laboratory, UK) has been created to allow secure exchange of information between sites, including SPINE nodes. Following test exchanges, the first exchange of real data in September 2005 transferred crystal-shipping information between Oxford and the UK-funded beamline at the ESRF, BM14. This has been followed by similar transactions involving Oxford, York, one further academic site and one industrial user. The same information-interchange standards are being adopted at other European synchrotrons. Although agreements on standards can be difficult to achieve, especially for standards as complex as data models, progress is being made and real benefits are in sight. These advances have depended to a large extent on the impetus given to collaboration by projects such as SPINE.

2.5. Crystallization and data tools

The Oxford partners have integrated the crystallization process into their protein production LIMS, Nautilus. The LIMS is used to record the details of all proteins going into the crystallization facility and stores the layout of the standard and optimization crystallization screens. The liquid-handling robots used in the set-up of crystallization trials have been interfaced to Nautilus *via* custom-written control software. Custom-written graphical user interfaces and the use of barcode scanners keep manual data entry to a minimum. Once the crystallization trial has been set up, the various data-management aspects of imaging and crystal identification are managed through the *PlateDB* and *Vault* software (Mayo *et al.*, 2005; see also Berry *et al.*, 2006). The system automatically detects and processes new images of the crystallization drops, recording information in a PostgreSQL database. All images acquired in Oxford are automatically evaluated on a local small Linux-based cluster using the *ALICE* software developed by the York partner (Wilson, 2002, 2004; Berry *et al.*, 2006). A web-based image-viewing facility allows users to monitor and annotate their experiments remotely. This system is freely available to academic laboratories and is being incorporated into PIMS to encourage uptake. Access to a demonstration account on the system is available at http://www.oppf.ox.ac.uk/vault/demo.php. Currently, 210 users have taken 31.6 million images from 13 500 plates created for 857 projects. The value of this annotated database is exemplified by the design of a 96-condition screen optimized for SPINE viral and human protein targets (see Berry *et al.*, 2006).

2.6. Structure determination

Software developments within SPINE have addressed structure determination by both NMR and crystallography. The NMR developments are detailed in AB *et al.* (2006) and we here concentrate on those for crystallography. SPINE partners, in collaboration with the BioXHIT project, have worked on the automated model-building package *ARP/wARP* and by utilizing structural bioinformatics to feed data into newly developed pattern-recognition software have extended the useful resolution range of this software (Morris *et al.*, 2004). Thus, recent versions of *ARP/wARP* can deal with lower resolution data and produce more complete models. These improvements were driven by a careful analysis of a limited number of experimental electron-density maps to improve the 'prior knowledge' of protein structure (Morris *et al.*, 2004; Cohen *et al.*, 2004). A more comprehensive statistical analysis is in progress using a larger number of data sets, collected in a database developed as a SPINE workpackage 7 initiative. This database captures, along with the final PDB model of a structure, additional crucial information which is not available from the PDB; namely, the experimental phase set derived from MIR/MAD/SAD experiments or molecular replacement. This database can be accessed at http://xtal.nki.nl/Depot and an interface for data searching and retrieval will be made available to other software method developers. All deposited data are stored in a relational database and submitted for curation and validation before being made available for statistical analysis. During the curation step, the data are standardized so that each data set can be run in the different analysis stages as simply as possible. The development of standard procedures for this task relates to the use of automated structure-solution pipelines developed within SPINE (see Bahar *et al.*, 2006). A long-term objective is to capture all SPINE-related structure solutions in this database.

2.7. Structure annotation

The ProFunc annotation server (Laskowski, Watson *et al.*, 2005; http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/) has been developed to help identify the likely biochemical function of a protein from its three-dimensional structure. ProFunc makes use of both existing and novel methods to analyse protein sequence and structure, identifying functional motifs or close relationships to functionally characterized proteins including, for example, a full analysis against the latest release of the Superfamily database. Often, where one method fails to provide any functional insight, another may be more helpful. A single-page summary of the analyses provides an at-a-glance view of what each of the different methods has found, whereas more detailed results are available on separate pages. The system has been made available to SPINE members and a large amount of testing has been performed, leading to the inclusion of new algorithms for finding structural motifs and similarities, better default parameters for many of the individual tasks and a clearer graphical presentation of the results. The system has been extended to ligand annotation since ligand prediction will play a crucial role in elucidating function from structure. Additional annotation is available through the new PDBsum service, http://www.ebi.ac.uk/thornton-srv/databases/pdbsum, where structure files can be uploaded and pages are produced providing at-a-glance views of the structure plus detailed structural analyses of each protein chain, any bound ligands and metals (Laskowski, Chistyakov *et al.*, 2005).

A second approach to docking aimed at high-throughput virtual screening for protein-function prediction has been developed at the EBI (Morris, 2004; Morris, Najmanovich *et al.*, 2005). This method aims at narrowing down ligand space by excluding those ligands whose shape does not match the predicted binding pocket. A new protein structure is analyzed with the program *SURFNET* (Laskowski, 1995) which picks up potential binding pockets by identifying surface clefts and indentations. Residue-conservation scores are computed with a maximum-likelihood phylogenetic approach and employed to reduce the size and enhance the quality of predicted binding pockets. Expansion is performed to describe the resulting three-dimensional shapes of the predicted binding pockets. Heuristics have been developed for three-dimensional object registration based on the first three moments of the Cartesian atomic coordinates, thus enabling the expansion coefficients to be used directly for shape comparison. This results in an extremely fast comparison metric based on the Euclidean distance in coefficient space.

3. Conclusions

Major developments in software to aid target identification and construct design have been made by several SPINE partners. Downstream, the significant progress made by SPINE partners in developing high-throughput procedures for protein expression, purification, crystallization, NMR and synchrotron data collection has stimulated the development of LIMS and informatics platforms crucial to managing the increased flow of data. Substantial progress has been made in NMR methods, for instance inferential structure determination (Rieping *et al.*, 2005), but further progress in X-ray data-processing, automated chain-tracing and refinement methods is required to streamline the stages in a structure solution that occur between mounting a crystal and obtaining a refined structure. Progress on increasing the functionality of the automatic density-interpretation program *ARP*/*wARP* is an example where SPINE has contributed. In this area of software, as in others, SPINE has played a valuable role in bringing together developers and users. SPINE has also addressed the relationship between protein structure and function (Goldsmith-Fischman & Honig, 2003) and whether the structure alone is sufficient to determine the biochemical function. While acknowledging that elucidating the biological role of a protein through structure is far from simple, progress has been made in automatic pipelines to annotate protein function from structure. The EBI has played a federating role in these developments by providing a central website and an automatic data-deposition system. This integrated site, containing details of the new experimental procedures, together with the data mined by the sequence and structure-annotation systems, provides a unique information resource for all the SPINE partners as well as the wider structural proteomics community.

The SPINE project was funded by the European Commission, contract No. QLG2-CT-2002-00988, under the RTD programme 'Quality of Life and Management of Living Resources'. The development of the ProFunc server was partly funded by the National Institutes of Heath, grant No. GM62414, the US DoE under contract W-31-109-Eng-38 and the BioSapiens Network of Excellence, funded by the European Commission, LSHG-CT-2003-503265.

References

AB, E. *et al.* (2006). *Acta Cryst.* D**62**, 1150–1161.

Alexandrov, N. N., Nussinov, R. & Zimmer, R. M. (1995). *Pacific Symposium on Biocomputing '96*, edited by L. Hunter & T. E. Klein, pp. 53–72. Singapore: World Scientific Publishing.

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.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2004). *Nucleic Acids Res.* **32**, D226–D229.

Andreini, C., Bertini, I. & Rosato, A. (2004). *Bioinformatics*, **20**, 1373–1380.

Au, K. *et al.* (2006). *Acta Cryst.* D**62**, 1267–1275.

Bahar, M. *et al.* (2006). *Acta Cryst.* D**62**, 1170–1183.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. & Yeh, L. S. (2005). *Nucleic Acids Res.* **33**, D154–D159.

Banci, L. & Rosato, A. (2003). *Acc. Chem. Res.* **36**, 215–221.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. & Eddy, S. R. (2004). *Nucleic Acids Res.* **32**, D138–D141.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. (2005). *Nucleic Acids Res.* **33**, D34–D38.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

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.

Bertini, I. & Rosato, A. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 3601–3604.

Beteva, A. *et al.* (2006). *Acta Cryst.* D**62**, 1162–1169.

Bianchetti, L., Thompson, J. D., Lecompte, O., Plewniak, F. & Poch, O. (2005). *J. Bioinform. Comput. Biol.* **3**, 929–947.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. (2003). *Nucleic Acids Res.* **31**, 365–370.

Boutselakis, H. *et al.* (2003). *Nucleic Acids Res.* **31**, 458–462.

Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999). *Nature Genet.* **23**, 151–157.

Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B. & Mornon, J. P. (1997). *Cell. Mol. Life Sci.* **53**, 621–645.

Castagnetto, J. M., Hennessy, S. W., Roberts, V. A., Getzoff, E. D., Tainer, J. A. & Piquet, M. E. (2002). *Nucleic Acids Res.* **30**, 379–382.

Chen, L., Oughtred, R., Berman, H. M. & Westbrook, J. (2004). *Bioinformatics*, **20**, 2860–2862.

Cipriani, F. *et al.* (2006). *Acta Cryst.* D**62**, 1251–1259.

Cohen, S. X., Morris, R. J., Fernandez, F. J., Ben Jelloul, M., Kakaris, M., Parthasarathy, V., Lamzin, V. S., Kleywegt, G. J. & Perrakis, A. (2004). *Acta Cryst.* D**60**, 2222–2229.

Coutinho, P. M. & Henrissat, B. (1999). *Recent Advances in Carbohydrate Bioengineering*, edited by H. J. Gilbert, G. Davies, B. Henrissat & B. Svensson, pp. 3–12. Cambridge: The Royal Society of Chemistry.

DuBois, P. (1999). *MySQL.* Indianapolis, IN, USA: New Riders Publishing.

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.

Esnouf, R. M., Hamer, R., Sussman, J. L., Silman, I., Trudgian, D., Yang, Z.-R. & Prilusky, J. (2006). *Acta Cryst.* D**62**, 1260–1266.

Ferron, F., Rancurel, C., Longhi, S., Cambillau, C., Henrissat, B. & Canard, B. (2005). *J. Gen. Virol.* **86**, 743–749.

Fulton, K. F., Ervine, S., Faux, N., Forster, R., Jodun, R. A., Ly, W., Robilliard, L., Sinsini, J., Whelan, D., Whisstock, J. C. & Buckle, A. M. (2004). *Acta Cryst.* D**60**, 1691–1693.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. & Bairoch, A. (2005). *The Proteomics Protocols Handbook*, edited by J. M. Walker, pp. 571–607. Totowa, NJ, USA: Humana Press.

Goh, C. S., Lan, N., Echols, N., Douglas, S. M., Milburn, D., Bertone, P., Xiao, R., Ma, L. C., Zheng, D., Wunderlich, Z., Acton, T., Montelione, G. T. & Gerstein, M. (2003). *Nucleic Acids Res.* **31**, 2833–2838.

Goldsmith-Fischman, S. & Honig, B. (2003). *Protein Sci.* **12**, 1813–1821.

Haebel, P. W., Arcus, V. L., Baker, E. N. & Metcalf, P. (2001). *Acta Cryst.* D**57**, 1341–1343.

Hamm, G. H. & Cameron, G. N. (1986). *Nucleic Acids Res.* **14**, 5–10.

Harris, M. & Jones, T. A. (2002). *Acta Cryst.* D**58**, 1889–1891.

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W. & Godzik, A. (2005). *Nucleic Acids Res.* **33**, W284–W288.

Junier, T. & Bucher, P. (1998). *In Silico Biol.* **1**, 13–20.

Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. E. (2000). *J. Mol. Biol.* **299**, 499–520.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). *J. Mol. Biol.* **235**, 1501–1531.

Kyte, J. & Doolittle, R. F. (1982). *J. Mol. Biol.* **157**, 105–132.

Laskowski, R. A. (1995). *J. Mol. Graph.* **13**, 323–330.

Laskowski, R. A., Chistyakov, V. & Thornton, J. M. (2005). *Nucleic Acids Res.* **33**, D266–D268.

Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005). *Nucleic Acids Res.* **33**, W89–W93.

Macleod, M. R. (2002). *J. Neurol. Neurosurg. Psychiatry*, **73**, 746.

Mar Albà, M., Lee, D., Pearl, F. M. G., Shepherd, A. J., Martin, N., Orengo, C. A. & Kellam, P. (2001). *Nucleic Acids Res.* **29**, 133–136.

Marchler-Bauer, A. *et al.* (2005). *Nucleic Acids Res.* **33**, D192–D196.

Matthew, N. & Stones, R. (2001). *Beginning Databases with PostgreSQL.* Indianapolis, IN, USA: Wrox Press.

Mayo, C. J., Diprose, J. M., Walter, T. S., Berry, I. M., Wilson, J., Owens, R. J., Jones, E. Y., Harlos, K., Stuart, D. I. & Esnouf, R. M. (2005). *Structure*, **13**, 175–182.

Morris, C., Wood, P., Griffiths, S., Wilson, K. S. & Ashton, A. W. (2005). *Proteins*, **58**, 285–289.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K. & Olson, A. J. (1998). *J. Comput. Chem.* **19**, 1639–1662.

Morris, R. J. (2004). *Acta Cryst.* D**60**, 2133–2143.

Morris, R. J., Najmanovich, R., Kahraman, A. & Thornton, J. M. (2005). *Bioinformatics*, **21,** 2347–2355.

Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A. & Lamzin, V. S. (2004). *J. Synchrotron Rad.* **11**, 56–59.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A. & Li, P. (2004). *Bioinformatics*, **20**, 3045–3054.

Page, R., Grzechnik, J. M., Spraggon, G., Kreusch, A., Kuhn, P., Stevens, R. C. & Lesley, S. A. (2003). *Acta Cryst.* D**59**, 1028–1037.

Pajon, A. *et al.* (2005). *Proteins*, **58**, 278–284.

Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., Mochel, T., Moulinier, L., Muller, A., Muller, J., Prigent, V., Ripp, R., Thierry, J.-C., Thompson, J. D., Wicker, J. & Poch, O. (2003). *Nucleic Acids Res.* **31**, 3829–3832.

Prilusky, J., Oueillet, E., Ulryck, N., Pajon, A., Bernauer, J., Krimm, I., Quevillon-Cheruel, S., Leulliot, N., Graille, M., Liger, D., Tresaugues, L., Sussman, J. L., Janin, J., van Tilbeurgh, H. & Poupon, A. (2005). *Acta Cryst.* D**61**, 671–678.

Raymond, S., O'Toole, N. O. & Cygler, M. (2004). *Proteome Sci.* **2**, 4.

Reinhardt, A. & Eisenberg, D. (2004). *Proteins*, **56**, 528–538.

Rieping, W., Habeck, M. & Nilges, M. (2005). *Science*, **309**, 303–306.

Romero, P., Obradovic, Z., Li, X., Garner, E., Brown, C. & Dunker, A. K. (2001). *Proteins*, **42**, 38–48.

Romier *et al.* (2006). *Acta Cryst.* D**62**, 1232–1242.

Rost, B. (1996). *Methods Enzymol.* **266**, 525–539.

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.

Sonnhammer, E. L. L., von Heijne, G. & Krogh, A. (1998). *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, edteid by J. Glasgow, T. Littlejohn, F. Major, R. Lathrop, D. Sankoff & C. Sensen, pp. 175–182. Menlo Park, CA, USA: AAAI Press.

Stevens, R. C. (2004). *Nature Struct. Mol. Biol.* **11**, 293–295.

Stevens, R. C., Yokoyama, S. & Wilson, I. A. (2001). *Science*, **294**, 89–92.

The Gene Ontology Consortium (2001). *Genome Res.* **11**, 1425–1433.

Thompson, J. D., Plewniak, F., Ripp, R., Thierry, J. C. & Poch, O. (2001). *J. Mol. Biol.* **314**, 937–951.

Thompson, J. D., Prigent, V. & Poch, O. (2004). *Nucleic Acids Res.* **32**, 1298–1307.

Thompson, J. D., Thierry, J. C. & Poch, O. (2003). *Bioinformatics*, **19.** 1155–1161.

Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L. & Rapp, B. A. (2001). *Nucleic Acids Res.* **29**, 11–16.

Wilson, J. (2002). *Acta Cryst.* D**58**, 1907–1914.

Wilson, J. (2004). *Crystallogr. Rev.* **10**, 73–84.

Zandstra, M. (2004). *PHP5 Objects, Patterns and Practice.* Berkeley, CA, USA: Apress.

Zhang, Z., Schaffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V. & Altschul, S. F. (1998). *Nucleic Acids Res.* **26**, 3986–3990.

Zolnai, Z., Lee, P. T., Li, J., Chapman, M. R., Newman, C. S., Phillips, G. N. Jr, Rayment, I., Ulrich, E. L., Volkman, B. F. & Markley, J. L. (2003) *J. Struct. Funct. Genomics*, **4**, 11–23.