3 Biotech (2012) 2:89–113
DOI 10.1007/s13205-011-0037-1

REVIEW ARTICLE

Current methods in structural proteomics and its applications in biological sciences

Babu A. Manjasetty · Konrad Büssow · Santosh Panjikar · Andrew P. Turnbull

Received: 6 October 2011 / Accepted: 9 November 2011 / Published online: 10 December 2011

Abstract A broad working definition of structural proteomics (SP) is that it is the process of the high-throughput characterization of the three-dimensional structures of biological macromolecules. Recently, the process for protein structure determination has become highly automated and SP platforms have been established around the globe, utilizing X-ray crystallography as a tool. Although protein structures often provide clues about the biological function of a target, once the three-dimensional structures have been determined, bioinformatics and proteomics-driven strategies can be employed to derive their biological activities and physiological roles. This article reviews the current status of SP methods for the structure determination pipeline, including target selection, isolation, expression, purification, crystallization, diffraction data collection, structure solution, refinement and functional annotation.

B. A. Manjasetty (✉)
European Molecular Biology Laboratory,
Grenoble Outstation and Unit of Virus Host-Cell Interactions,
UJF-EMBL-CNRS, UMI 3265, 6 rue Jules Horowitz,
BP181, 38042 Grenoble Cedex 9, France
e-mail: babu@embl.fr

K. Büssow
Department of Molecular Structural Biology,
Helmholtz Centre for Infection Research,
38124 Braunschweig, Germany

S. Panjikar
Australian Synchrotron, 800 Blackburn Road,
Clayton, VIC 3168, Australia

A. P. Turnbull
Cancer Research Technology Ltd., Birkbeck College,
University of London, London WC1E 7HX, UK

Keywords Protein structure analysis · X-ray crystallography · Bioinformatics · Structural proteomics

Introduction

One of the most spectacular recent achievements in life sciences has been the sequencing of the entire human genome, accomplished by the Human Genome Project. The resolution of the entire sequence of the human genome has resolved many unanswered questions relating to human life. The human body comprises a vast number of cells and each cell contains many thousands of different proteins necessary to maintain cellular function. Knowledge of the sequence of the human genome means that disease-associated abnormalities can now be detected at the genetic level. Furthermore, sequence comparisons can provide an insight into the evolutionary relationship between organisms. As of August 2011, the UniProtKB/Swiss-Prot database has contained in excess of half a million non-redundant sequence entries. Hence, it is clear that large-scale genomic projects have provided the sequence infrastructure for the in-depth analysis of proteins. A new fundamental concept of the proteome (PROTEin complement to a genOME) has emerged that aims to unravel the biochemical and physiological mechanisms of complex multivariate diseases at the functional and molecular level. As a consequence, the new science of proteomics has been established to complement physical genomic research. Proteomics can be defined as the *qualitative* and *quantitative* comparison of proteomes under different conditions, which aims to further characterize biological processes and functional protein networks (Naistat and Leblanc 2004; Petschnigg et al. 2011; Stults and Arnott 2005). However, the knowledge gleaned from the various genomes

sequenced to date is not sufficient to understand the function of proteins within the cell. To characterize functional protein networks and their dynamic alteration during physiological and pathological processes, proteins have to be identified, sequenced, categorized and classified with respect to their function and interaction partners. To understand their functions at a molecular level, it is often necessary to determine their three-dimensional (3D) structures at atomic resolution.

During the past decade, the emerging field of structural proteomics (SP) has developed, representing an international effort aimed at the large-scale determination of the 3D structures of proteins encoded by the genomes of key organisms (Burley 2000; Joachimiak 2009; Manjasetty et al. 2007; Terwilliger 2011). Initiatives in SP research have led to the development of novel strategies and automated protein structure determination pipelines around the world (Table 1) (Chance et al. 2004; Manjasetty et al. 2008).

When protein structure analysis was first established in the late 1960s and the X-ray structures of myoglobin and hemoglobin were determined, the development of such a high-throughput (HT) infrastructure for protein structure analysis would have seemed like an impossible dream. The remarkable success and technological advancements since then have had a tremendous impact on throughput in protein structure determination and all stages of the pipeline have become more or less automated (Fig. 1). Currently, SP initiatives are generating protein structures at an unprecedented rate and have resulted in an exponential growth in the number of protein structures deposited in the Protein Data Bank (Fig. 2: 65979 PDB entries, as of August 2011). However, the number of solved protein structures in the PDB represents only a small proportion of the theoretical number of proteins encoded by genomic sequences.

To bridge this gap and to meet the demand of rapidly obtaining protein structure information, advancements have been made in SP methodologies in the form of HT

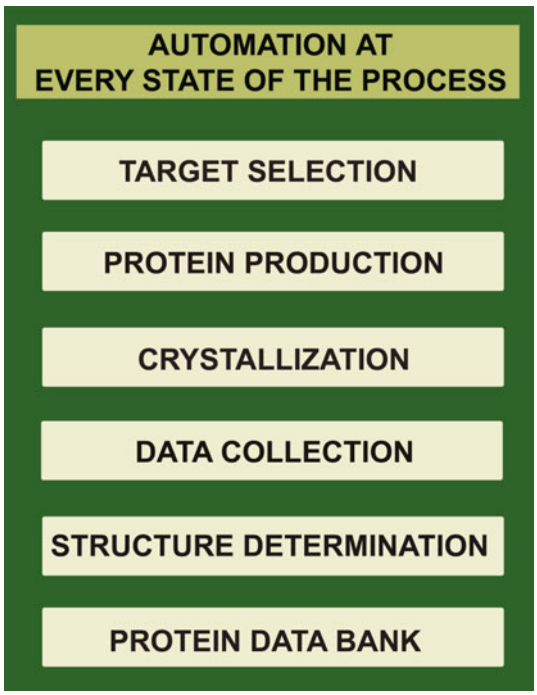

Fig. 1 Process involved in SP using X-ray crystallography

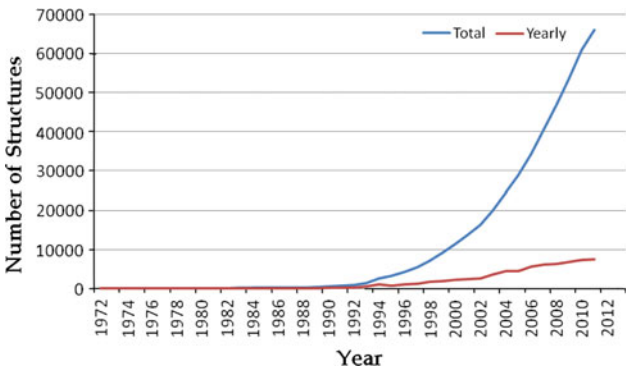

Fig. 2 Exponential growth in the number of X-ray protein structures deposited in the Protein Data Bank

Table 1 Major centers for high-throughput structure determination around the world

No.	Country	Center	Web address	PDB entries
1.	Japan	RIKEN Structural Genomics/Proteomics Initiative	http://www.rsgi.riken.go.jp/	2,702
2.	USA	Midwest Center for Structural Genomics (MCSG)	http://mcsg.anl.gov/	1,389
3.	USA	Joint Center for Structural Genomics (JCSG)	http://www.jcsg.org/	1,234
4.	USA	New York Structural Genomics Research Consortium (NYSGRC)	http://www.nysgrc.org/	1,028
5.	Canada, UK, Sweden	Structural Genomics Consortium (SGC)	http://www.thesgc.org/	1,005
6.	USA	Northeast Structural Genomics Consortium (NESG)	http://www.nesg.org/	964

technologies. However, these technologies have encountered some of the traditional bottlenecks in structure determination for difficult proteins and complexes of proteins at HT. To overcome these bottlenecks, efforts have been focused on improving the structure determination pipeline by streamlining and optimizing protein production, protein crystallization, data collection and structure solution. In addition, SP centers have adopted bioinformatics analysis of potential targets to generate models based on solved structures and to establish collaborative research to exploit the function of proteins. Recently, in the USA, the National Institutes of Health established a *Protein Structure Initiative* (PSI): a biology network to determine protein structures including membrane proteins of high biological interest. The objective of the PSI is to develop suitable technologies for membrane protein structure solution, using bioinformatics and modeling to leverage solved structures, and to carry out collaborative research to provide a link between a structure and its biomedical and biotechnological impact. On the other hand, in Europe, the emphasis for the *Structural Proteomics IN Europe* (SPINE) initiative has been to apply these HT technologies to systems of biological interest, the ultimate aim being to solve significant biological problems more effectively. Furthermore, the European *INSTRUCT* project offers scientists access to world-class structural biology and SP infrastructures and expertise. *INSTRUCT* makes integration possible more rapidly, creating a coherent forum for structural biology. This forum will stimulate closer collaboration between scientific communities and initiatives in biological sciences.

In this report, recent advances in protein structure analysis in the context of SP will be discussed. Furthermore, the impact of SP on other biological sciences including drug discovery and biotechnology will be explored.

Automation and strategies for protein structure analysis

Protein production and crystallization

Generating pure, soluble and homogeneous protein for structure determination is a major rate-limiting step in the overall process. Traditional sequential generation of single expression constructs for a single protein target has been superseded by parallel, HT cloning techniques. Genetic engineering and the use of specific crystallization chaperones are two approaches that have proven invaluable for the determination of many highly important protein structures.

Screening of candidate proteins

Structural biology projects are typically initiated to characterize the biological activity of a specific protein or protein complex. In some cases, crystallization of that exact protein leads to a structure that can be correlated directly to functional data. However, in many cases, researchers will eventually come to the conclusion that the protein of interest is not suitable for structural analysis. It makes sense, therefore, to include parallel, HT approaches early on for identifying optimal boundaries and experimental conditions for protein production and crystallization. The selection of candidate proteins is very much project dependent, but will usually include orthologs or homologs of the original protein of interest and genetic constructs corresponding to subregions or individual domains. Methods for HT characterization of larger numbers of expression clones have originally been described for bacterial expression systems (Berrow et al. 2006; Büssow et al. 2005). Small-scale expression testing is more difficult to achieve in eukaryotic systems such as yeast (Holz et al. 2003) and baculovirus (McCall et al. 2005). *Transient transfection* of mammalian cell lines such as HEK293 is a highly efficient system for secreting mammalian glycoproteins and has also been successfully applied to produce membrane proteins such as rhodopsin (Standfuss et al. 2011). This method can be performed in a HT manner for characterization of protein candidates for crystallization (Lee et al. 2009).

Glycoproteins

The choice of expression system has a great influence on the quality and quantity of the produced recombinant protein. *Cell-free* protein production has proven its value for producing soluble (Makino et al. 2010) and membrane proteins (Junge et al. 2011; Reckel et al. 2010) for NMR and crystallographic studies (Watanabe et al. 2010). Mammalian proteins stabilized by disulfide bonds and modified by glycosylation are especially demanding targets. Mammalian cells are the ideal host for these proteins, since they yield protein with all the post-translational modifications required for biological activity, including authentic glycosylation and correct disulfide pairing. However, cell culture is time and labor intensive. Many extracellular mammalian proteins can be recovered in active form through refolding of bacterial inclusion body proteins (Vallejo and Rinas 2004). Obviously, these proteins do not require post-translational modifications other than disulfide bridges for correct folding. *Refolding* of proteins from inclusion bodies is common in industrial production but requires extensive process optimization. Structural biology projects applying inclusion body

refolding benefit from automated screening of folding conditions with generic, biophysical assays (Cowieson et al. 2006; Scheich et al. 2004; Vincentelli et al. 2004).

Animal cell lines are highly effective for the secretion of proteins with native glycosylation and disulfide bonds. Glycoproteins produced with the mammalian CHO or HEK293 cell lines carry heterogeneous, complex-type oligosaccharide chains attached to Ser/Thr (O-linked) or Asn (N-linked) side chains. Crystallization of glycoproteins is difficult because of the heterogeneity and flexible conformation of the bulky oligosaccharides, which can also mask possible sites of crystal contacts on the protein surface. Some glycosylation sites can be removed by mutagenesis. Regions with O-linked glycosylation are generally proline-rich and unfolded, and can be excluded from genetic constructs. However, many proteins require glycosylation for folding and transport through the secretory pathway. Enzymatic removal of N-linked glycans from the purified protein with endoglycosidase H or F leaves a single monosaccharide attached, which may increase the solubility of the deglycosylated protein. Enzymatic deglycosylation is efficient for oligosaccharides of the high-mannose type as obtained from the baculovirus system (Fig. 3a). Processing of N-linked glycans by mammalian cell lines results in complex-type oligosaccharides that are difficult to cleave enzymatically. Complex-type glycosylation can be prevented by chemical glycosylation inhibitors (Chang et al. 2007) or by mutating the host cells. The gene for the enzyme *N*-acetylglucosaminyl-transferase I (GnTI), which modifies high-mannose type oligosaccharides, has been mutated in the cell lines CHO Lec1, Lec3.2.8.1 (Stanley 1989) and HEK293S-GnTI$^{(-)}$ (Reeves et al. 2002). These cell lines and normal HEK293 cells treated with the glycosylation inhibitors kifunensine or swainsonine have enabled the production of many glycoproteins and their crystallization upon enzymatic deglycosylation (Aricescu et al. 2006; Chang et al. 2007; Davis et al. 1993; Standfuss et al. 2011). Optimized protocols and cell lines allow performing transient transfection of HEK293 at up to liter scale with inexpensive reagents (Aricescu et al. 2006). However, not all proteins can be produced in sufficient amounts by transient transfections. Stable cell lines allow the production of proteins more reproducibly and in much larger volumes in bioreactors. However, establishing lines with good performance requires considerable effort. Novel approaches of *stable cell line* development, based on preparative cell sorting and *recombinase-mediated cassette exchange* (*RMCE* Fig. 3c), combine faster development times with improved performance and have been used successfully for X-ray crystallography studies (Wilke et al. 2010, 2011).

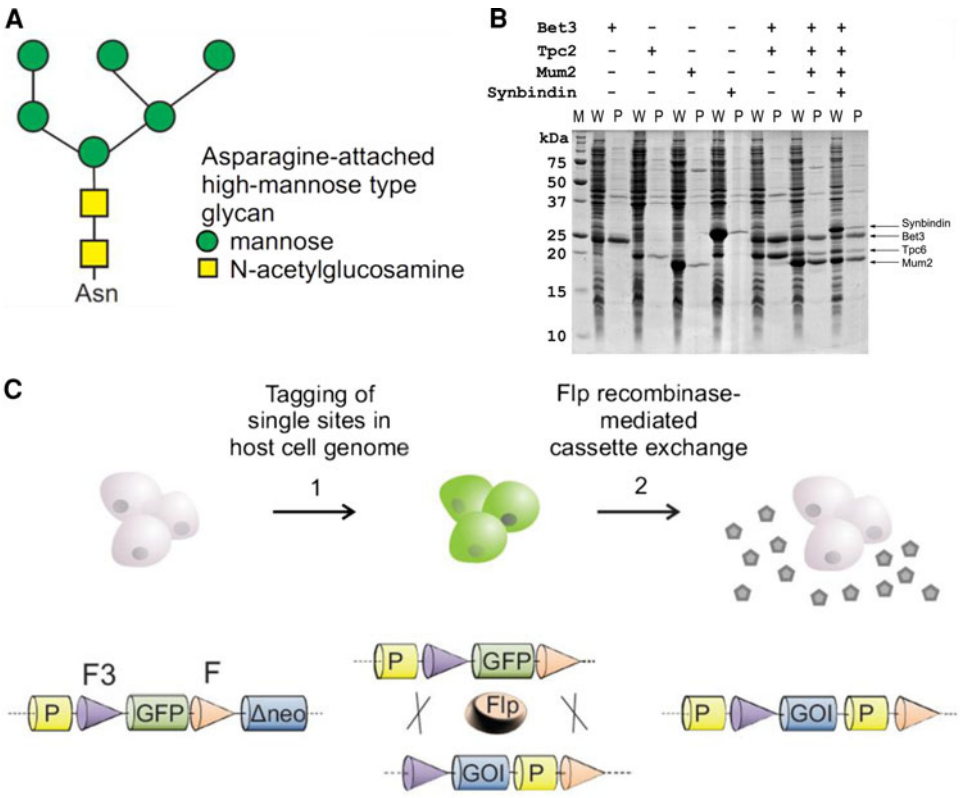

Fig. 3 Protein production. **a** Glycosylation: structure of a high mannose-type glycan. **b** Co-expression of a complex of four proteins with the pQLink system. *M*: marker, *W*: whole cellular protein, *P*: purified protein (Scheich et al. 2007). **c** Cell line development by *recombinase-mediated cassette exchange* (RMCE): cells are transfected with a vector containing a *GFP* gene flanked by recombination sites *F3* and *F* and GFP-positive cells are isolated. Cassette exchange is initiated by co-transfecting a tagged cell line with an Flp recombinase expression vector and a targeting vector bearing the gene of interest (GOI). Flp recombinase exchanges the tagging gene cassette and a production cell line is obtained (Wilke et al. 2011)

Crystallization chaperones

Some protein families require a combination of specialized strategies for successful crystallization. Crystallization chaperones are proteins that specifically bind to the target protein and support "carrier-driven" crystal growth (reviewed by (Koide 2009)). They limit the conformational flexibility of the target protein and provide a large, hydrophilic interaction surface for initiating crystal lattice contacts. *Fab fragments* of monoclonal antibodies have been used traditionally as crystallization chaperones. In addition, recombinant antibodies from camels (V_HHs, also called "nanobodies") and synthetic scaffold proteins have demonstrated their usefulness in many examples (Koide 2009). Disulfide-free synthetic scaffold proteins such as *designed ankyrin repeat proteins* (*DARPINs*) or the fibronectin type III domain (FN3) can be screened for specific binders in vitro and can be produced easily in *E. coli*. Fab fragments enabled the first crystal structure to be determined for a non-rhodopsin GPCR (Rasmussen et al. 2007) and a full-length potassium channel (Uysal et al. 2009). Furthermore, the first high-resolution crystal structure of the β2 adrenergic receptor–Gs protein complex has recently been reported in which nanobodies (camelid antibody fragments) were used to significantly improve crystal quality (Rasmussen et al. 2011). Nanobodies are relatively simple proteins, about a tenth the size of antibodies and just a few nanometers in length.

Protein engineering

Protein engineering can overcome problems with producing sufficient amounts of protein, keeping protein soluble at the concentrations required for crystallization and obtaining proteins with surfaces that allow crystal formation (Derewenda 2010). In general, protein engineering follows one of three strategies: designing shortened proteins lacking terminal residues outside the globular fold, mutating residues on the target protein's surface or designing fusion proteins.

If a full-length protein cannot be produced or crystallized, then a common strategy is to design shorter variants which represent isolated domains, eliminating flexible regions at the termini or large internal loops. Databases such as *PFAM* provide information on the presence of *conserved domains* in protein sequences. Alternatively, genetic constructs can be designed that avoid regions predicted to be disordered and unfolded by software tools such as DISOPRED2 (Ward et al. 2004), RONN (Yang et al. 2005), or the meta-server metaPrDOS (Ishida and Kinoshita 2008). The strategy of designing genetic constructs on the basis of computational analysis may fail because of imprecise or missing information in the respective databases. Robotic screening of random truncation libraries represents an alternative technique in such cases (reviewed by (Dyson 2010; Yumerefendi et al. 2011)). In this technique, the cDNA is fragmented and a library of expression clones is created by cloning the fragments. By chance, a few clones of the library will contain a fragment that encodes for just one complete domain. Such clones will express soluble protein. Different ways of screening libraries for clones that express soluble protein have been described, including a filtration technique (Cornvik et al. 2005), the biotinylation assay of the *ESPRIT* technology (Yumerefendi et al. 2011) and screening based on GFP fusion protein fluorescence (Pedelacq et al. 2011). The *ESPRIT* (*Expression of Soluble Proteins by Random Incremental Truncation*) library technology has been adapted recently to allow screening for soluble protein complexes (An et al. 2011).

Single point mutations can have a dramatic effect on a protein's solubility and crystal formation (Derewenda 2010). The most successful point mutation strategy, *surface entropy reduction* (SER), replaces small clusters of two to three surface residues with high conformational entropy such as Lys, Glu or Gln with Ala. SER produces mutants that are often more susceptible to crystallization than the wild-type protein. More than 100 structures of proteins optimized by SER have been solved. A Web server facilitates protein engineering for SER (http://services.mbi.ucla.edu/SER/). In general, SER does not improve protein solubility. Proteins that cannot be concentrated to sufficient levels for crystal growth benefit from strategies opposite to SER. The solubility of such proteins can often be increased by reducing the hydrophobicity of surface residues. This approach is more difficult than SER, because exchanging hydrophobic surface residues requires some knowledge of the protein's structure.

Directed point mutations are generally not used to improve the protein production levels since no rational strategies are available. However, screening of large libraries of random mutants of the target proteins, enabled by laboratory automation, has been successful (Cornvik et al. 2005; Listwan et al. 2009; Yumerefendi et al. 2011). Fusion proteins with partners such as glutathione S-transferase (GST), thioredoxin or maltose binding protein (MBP) are a more common approach to improve the target protein's production and solubility. However, the flexible linker between the fusion partners generally inhibits crystallization. Furthermore, when the fusion partner is removed using a site-specific protease, the improvement in solubility conferred by the fusion partner may be lost. Careful design of MBP fusion proteins enables *carrier-driven crystallization* of intact fusion proteins (Moon et al. 2010). These fusions have to be designed in such a way that the MBP's C-terminal α-helix is fused directly to the globular core of the target protein, thereby avoiding

flexibility between the fusion partners. Then, the MBP part can improve protein yield and solubility and promote crystal growth.

One method of surface modification that does not involve additional cloning is reductive lysine methylation, where lysine side chains are chemically modified (Sledz et al. 2010; Walter et al. 2006). The technique can improve the X-ray diffraction of existing crystals, or permit the crystallization of proteins that had previously failed to yield crystals.

Protein complexes

Protein complexes are attractive targets for X-ray crystallography, because their structures reveal important information relating to the molecular details of specific protein recognition. However, crystallization of a complex requires careful preparation that includes critical assessment of the available data, careful optimization of sample preparation and functional and biophysical characterization of the complex using a variety of methods (Collinet et al. 2011; Perrakis et al. 2011). Very stable complexes that do not dissociate are preferred targets for crystallization. However, the subunits of such stable heterocomplexes may not be able to fold into a soluble conformation alone, necessitating the *co-expression* of the complex components. Transient complexes, on the other hand, which exist in equilibrium with the dissociated subunits, are more difficult to crystallize because of sample heterogeneity. Subunits of transient complexes may form crystals that exclude the other subunit, which is often difficult to detect. Recombinant production of the subunits of a protein complex in the same host cell by co-expression has been described with a large variety of systems (Busso et al. 2011; Nie et al. 2009; Vijayachandran et al. 2011). Novel cloning strategies enable co-expression of many subunits in host cells including *E. coli*, baculovirus and mammalian cells (Berger et al. 2004; Kriz et al. 2010; Trowitzsch et al. 2010), and have been adapted to automated cloning (Bieniossek et al. 2009). The *pQLink* system (Scheich et al. 2007) allows co-expression of an unlimited number of protein subunits in *E. coli* with different affinity tags (Fig. 3b). *pQLink* vectors have been widely used by different laboratories, mainly for eukaryotic vesicle tethering complexes (Kummel et al. 2008; Lees et al. 2010; Ren et al. 2009). Studies comparing a large variety of expression systems have demonstrated that subtle changes in the expression strategy have a profound effect on the success of co-expression experiments, even if the main parameters, protein sequence and host cell are identical (Busso et al. 2011).

Successful recombinant expression of protein complexes requires that the subunits are synthesised in similar amounts. Otherwise, the yield of the complete complex is determined by the subunit present in the lowest concentration. Also, a heterogeneous mixture of the complete complex with smaller oligomers not comprising all subunits is obtained. To circumvent this problem, the synthesis of *polyproteins* has been introduced for generating protein complexes (Vijayachandran et al. 2011). This strategy is reminiscent of the genomes of many viruses that contain large open reading frames encoding *polyproteins* that are cleaved by viral proteases into single proteins upon translation. A baculovirus vector containing a large open reading frame comprising single protein sequences separated by a site-specific protease site was created. The coding sequence of the TEV protease was included in the vector. Upon overexpression, intracellular TEV protease cleaved the polyprotein into single subunits of a protein complex. This strategy was successfully demonstrated for sub-complexes of human general transcription factor TFIID and other complexes (Vijayachandran et al. 2011).

Protein crystallization methods and automation

Production of protein crystals suitable for structural studies poses one of the major bottlenecks in the entire process. Finding crystallization conditions that yield single, well-ordered crystals with low mosaicity that diffract to sufficient resolution can be very challenging. The quality of a crystal is often linked to the number of crystals formed (a few large crystals versus many microcrystals), size (larger is better) and appearance (optically clear, sharply faceted crystals are best). However, any true measure of quality must verify that the diffraction properties correlate with the morphological quality of the crystal.

Crystallization can occur spontaneously, or alternatively it can take several days, weeks or months for crystals to appear. Longer crystallization times are usually indicative of proteolytic cleavage at the protein termini promoting crystal formation. It is not easy to provide an estimate for maximum protein crystal growing time. There have been some reports showing that, in some cases, protein crystals may take as long as 6 months or a year to appear. However, an average growing time for a protein crystal is typically less than a month. Normally, protein crystallization occurs when the concentration of protein in solution is greater than its limit of solubility, so that the protein solution becomes supersaturated. To crystallize a protein, it undergoes slow precipitation from an aqueous solution. As a result, individual protein molecules align themselves in a repeating series of "unit cells" by adopting a uniform orientation. One unavoidable aspect of crystallizing a newly expressed protein is the need to carry out a large number of experiments to find suitable conditions in which the protein crystallizes. It can be extremely tedious and time consuming to set up a broad

array of different crystallization experiments manually. With the advent of HT liquid handling and crystallization systems, it is relatively easy to prepare a thousand or more crystallization experiments in which crystallization parameters, such as the ionic strength, pH, protein and precipitant concentration and temperature, are varied systematically. However, the success rate does not depend upon the number of crystallization conditions tested.

Methods used for crystallization include vapor diffusion, batch crystallization, dialysis, seeding, free-interface diffusion and temperature-induced crystallization. The most popular method for setting up crystallization experiments is vapor diffusion, which includes hanging drop (for smaller volumes), sitting drop (for larger volumes), the sandwich drop, reverse vapor diffusion and pH gradient vapor diffusion methods. A drop containing a mixture of precipitant and protein solution is sealed in a chamber with pure precipitant. Water vapor subsequently diffuses from the drop until the osmolarity of the drop and the precipitant is equal. The dehydration of the drop causes a slow concentration change of both protein and precipitant until equilibrium is achieved, ideally in the crystal nucleation zone of the phase diagram (Dessau and Modis 2011). Batch crystallization relies on bringing the protein directly into the nucleation zone by mixing protein with the appropriate amount of precipitant. The batch method is usually carried out under oil to prevent the diffusion of water out of the drop (Chayen 1997). Many of these methods can be performed using HT automated instrumentation and miniaturization of crystallization experiments and have had huge impacts on protein crystallization in terms of saving time and conserving precious sample. For example, crystallization robots such as the *Phoenix*™ *RE* (Rigaku Corporation) and the *Mosquito*® (TTP Labtech), which can accurately and reproducibly dispense very small volumes (nl in size) into 96-well plates for automated screening and optimization of crystallization conditions, are now commonplace in many laboratories (Fig. 4). In addition, TTP

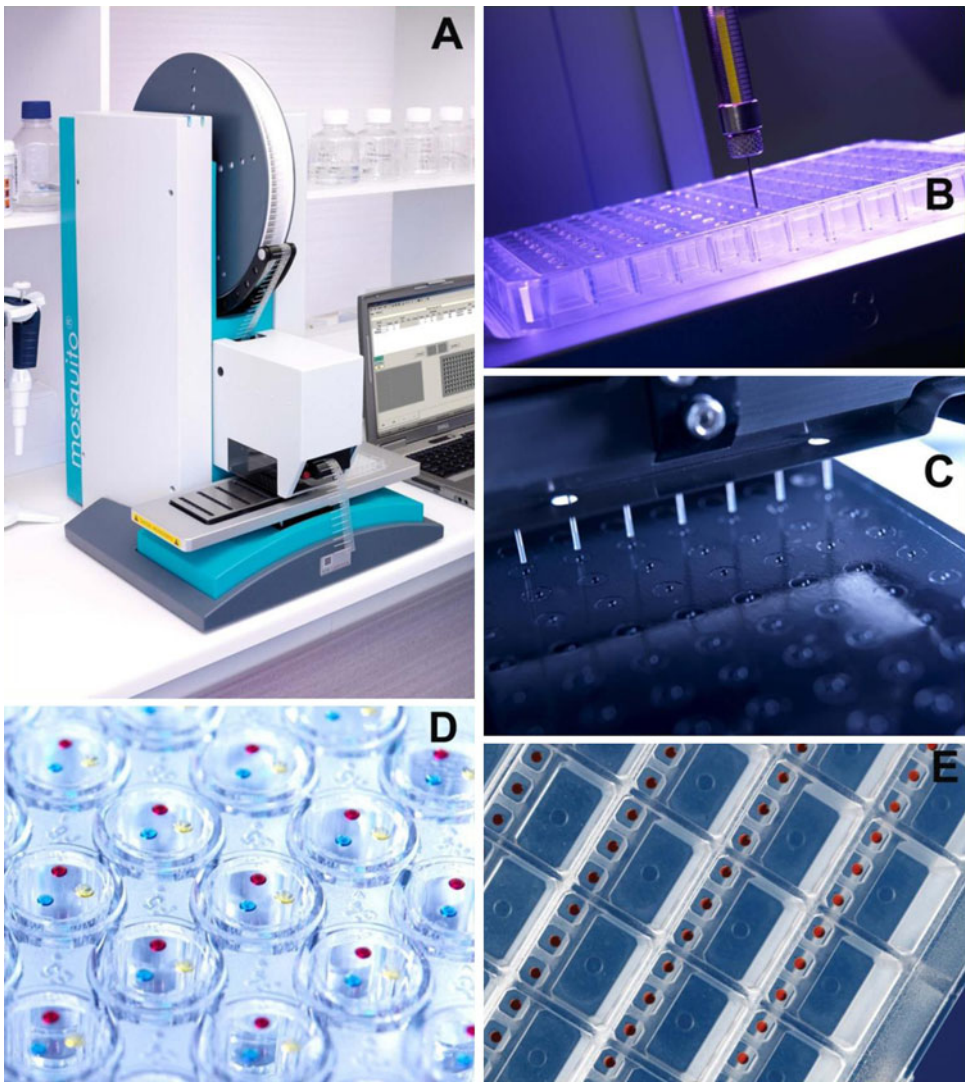

Fig. 4 Protein crystallization and automation. **a** TTP LabTech's mosquito® Crystal automates protein crystallography vapor diffusion set-ups, additive screening and microseeding; **b** TTP LabTech's mosquito® LCP: a dedicated instrument for crystallising membrane proteins using lipidic cubic phase screening. The panel highlights the positive displacement syringe, which dispenses the highly viscous lipid mesophases used in the LCP technique into 96-well crystallization plates. (**c** and **d**) Crystallization plate set up for hanging drop vapor diffusion experiments; **e** nanoliter sitting drop experiments set up in a 96-well plate. (*Images courtesy of TTP LabTech Ltd, UK*)

LabTech's *Mosquito® LCP* (Lipid Cubic Phase) has been designed to aid in the crystallization of membrane proteins by accurately dispensing nanoliter quantities of highly viscous lipids or detergents that are required to retain the structural integrity of the sample. A recent development in protein crystallization has been the use of high-density, chip-based microfluidic systems for crystallizing proteins using the free-interface diffusion method at nanoliter scale, including Emerald Biosystems *MPCS* (Microcapillary Protein Crystallization System)(Gerdts et al. 2008), Fluidigm Corporations *TOPAZ®* system (Segelke 2005) and the Microlytic *Crystal Former* (Stojanoff et al. 2011). These platforms have the advantage of using minimal protein sample to screen a broad range of crystallization conditions. The Rigaku CrystalMation™ system was set up to fully automate the crystallization process while dealing with sample volumes of 100 nl per experiment.

A popular strategy for the optimization of crystallization conditions in vapor diffusion is crystal seeding. Seeding decouples nucleation from crystal growth and involves transferring previously obtained seed crystals into undersaturated drops. Homogeneous seeding techniques include microseeding, streak seeding and macroseeding. Seed stock for microseeding can be conveniently generated using Hampton Research's Seed Bead kit. More recently, a simple, automated microseeding technique based on microseed matrix screening has been developed (D'Arcy et al. 2007). This method consists of the addition of seeds into the coarse screening procedure using a standard crystallization robot and has been shown to not only produce extra hits, but also generate better diffracting crystals. Successful cases for a simple semi-automated microseeding procedure for nanoliter crystallization experiments have also been recently described (Walter et al. 2008). Furthermore, crystallization plate storage and inspection are now fully automated. For example, the *Minstrel™ drop imager family* (Rigaku) and *Rock Imager* (Formulatrix) combine imagers with gallery plate hotels/incubators to store crystallization plates at a constant temperature, periodically inspect them and manage the data (Hiraki et al. 2006; Walter et al. 2005).

Despite the progress that has been made in increasing throughput, the act of identifying crystals in the crystallization experiments remains a task requiring human intervention. A number of attempts are being made to automate crystal detection from the imaged drop and varying degrees of success have been reported (Liu et al. 2008). Automated crystal recognition has the potential to reduce the time-consuming human effort for screening crystallization drop images. Several approaches have been suggested to increase contrast for imaging and detection of protein crystals in such cases: crystal birefringence (Echalier et al. 2004), addition of fluorescent dyes (Groves et al. 2007) and monitoring the fluorescence of trace labeled protein molecules (Forsythe et al. 2006). The identification of crystallization hits has been simplified by UV detection combined with conventional imaging (Judge et al. 2005). For example, the latest generation of imaging systems combine visible and UV inspections providing a powerful tool for monitoring crystallization trials: when crystals are still too small to be mounted, the intrinsic protein fluorescence signal gives confidence that a crystallization hit is worth pursuing. *Second-order nonlinear optical imaging of chiral crystals* (*SONICC*) is an emerging technique for crystal imaging and characterization (Kissick et al. 2010, 2011). SONICC imaging has been found to compare favorably with conventional optical imaging approaches for protein crystal detection, particularly in non-homogeneous environments that generally interfere with reliable crystal detection by conventional means.

A recent development is the *X-CHIP* (X-ray Crystallization High-throughput Integrated Platform): a novel microchip that provides a stable microbatch crystallization environment and combines multiple steps of the crystallographic pipeline from crystallization to diffraction data collection onto a single device (Kisselman et al. 2011). This system facilitates HT crystallization screening, visual crystal inspection, X-ray screening and data collection. The chip eliminates the need for manual crystal handling and cryoprotection of crystal samples while allowing data collection from multiple crystals in the same drop.

Data collection and processing

Data acquisition involves the recording of a series of X-ray diffraction images using a detector. The process of crystal mounting, centering, exposing with X-rays, recording diffraction data and dismounting the crystal represent the major steps in crystallographic data collection.

Radiation source, crystal handling and detector: past and present tools

In the past, protein crystals typically ranging in size from tenths of a millimeter to several millimeters were mounted in glass capillary tubes. To collect data, the capillary tube was mounted on a goniometer and exposed to X-rays at room temperature. These X-rays were generated by low flux, sealed tube sources. Nowadays, data collection is handled by automated sample changers and micro-diffractometers in a cryogenic (100 K) environment utilizing brighter synchrotron radiation as the X-ray source. Cryo-freezing the sample inhibits free radicals diffusing through the crystal during data collection: these free radicals cause secondary radiation damage that leads to degradation in the quality of collected data. There are currently in excess of

125 dedicated protein crystallography beamlines around the World. The X-ray films that were used for data recording in the past have now been superseded with charge-coupled devices (CCD) and pixel array detectors, which allow diffraction data to be recorded directly and stored straight to disk. For example, a recent development has been the *PILATUS* detector (*pixel apparatus for the SLS*), which has no readout noise, superior signal-to-noise ratio, a readout time of 5 ms and high dynamic range compared to CCD and imaging plate detectors. Delivery of high flux beam at third-generation synchrotron sources coupled with the advances in detector technology and control systems have significantly accelerated the speed of macromolecular diffraction data collection. An example of a state-of-the-art synchrotron X-ray data collection setup is shown in Fig. 5. Nowadays, crystals larger than 50 μm in size can be evaluated at conventional synchrotron beamlines. However, with some targets such as membrane proteins and multi-protein complexes, it is notoriously difficult to obtain crystals of sufficient size and order to generate high-quality diffraction data. Hence, next generation microfocus beamlines with reduced beam sizes have been established at synchrotron sites around the world, allowing measurements to be made on crystals a few micrometers in size. It has been predicted that a complete data set with a signal-to-noise ratio of 2σ at 2 Å resolution could be collectable from a perfect lysozyme crystal measuring just 1.2 μm in diameter using a microfocus beam (Holton and Frankel 2010). A number of crystal structures have been solved using micrometer-sized crystals by merging data from several crystals, including a polyhedron-like protein structure (~5–12 μm) (Coulibaly et al. 2007) and a thermally stabilized recombinant rhodopsin (with crystal dimensions of $5 \times 5 \times 90$ μm^3) (Standfuss et al. 2007). Recently, strategies have been developed to determine structures from showers of microcrystals using acoustic droplet ejection (ADE) to transfer 2.5 nl droplets from the surface of microcrystal slurries through the air and onto micromesh loops. Individual microcrystals are located by raster-scanning a several-micron X-ray beam across the cryocooled micromeshes. X-ray diffraction data sets are subsequently merged from several micrometer-sized crystals and this technique has been used to solve 1.8 Å resolution crystal structures (Soares et al. 2011).

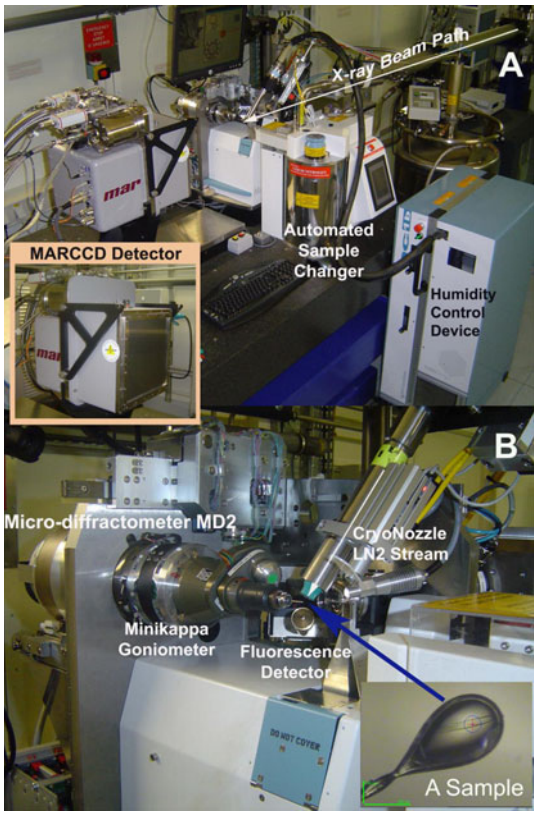

Fig. 5 X-ray data collection facility. **a** End-station instrumentation at ESRF beamline BM14 (http://www.bm14.eu) illustrating the sample changer used to exchange cryo-frozen crystals on the goniometer and the MARCCD (Marresearch GmbH) detector used to collect diffraction images. The *arrow* highlights the path of the X-ray beam. **b** Close-up view showing the frozen crystal sample in the center of the image and the surrounding beamline instrumentation. The *red cross* and *blue circle* represent the center and diameter of the X-ray beam on the frozen crystal sample (*bottom right*)

As a result of these technological advancements, the time required to setup a diffraction experiment has become a significant proportion of the total time of an experiment. The diffraction experiment involves sample mounting, crystal centering and determination of data collection parameters. Significant progress has been made in automating crystal mounting, crystal centering and the energy scan to find metals or ions present in crystals that can be used for phasing (Heinemann et al. 2003; Shi et al. 2005). Automated sample mounting systems allow users to mount samples on the beamline without entering the experimental hutch. These systems minimize the need for manual intervention and facilitate the rapid and systematic screening of dozens of samples. For example, the automated sample changers equipped at the EMBL/ESRF beamlines are capable of handling 50 frozen samples, whereas the *ACTOR* (*Automated Crystal Transfer, Orientation and Retrieval*) robots installed on the beamlines at the DIAMOND synchrotron can mount up to 80 cryogenically frozen samples from their onboard storage dewars. This facilitates the rapid screening and ranking of crystals and enables users to collect data from their best diffracting crystal(s) (Beteva et al. 2006; Cipriani et al. 2006). These features make the automated approach far quicker than manual operations insuring that beamtime is used efficiently.

Before starting data collection, the crystal needs to be aligned so that it is coincident with the X-ray beam and the

rotation axis. This is normally performed manually by the user at the beamline. However, for fully automated operation of the beamline, automated crystal centering is a prerequisite, especially when sample mounting robots are used. Semi-automated crystal centering based on a user clicking a mouse to indicate the position of the crystal through a specially designed software interface has been shown to be relatively robust and is employed at most synchrotron beamlines (Snell et al. 2004). Recent reports show that it is possible to center crystals automatically without user intervention using the recognition software *C3D* (Lavault et al. 2006), *XREC* (Pothineni et al. 2006) or alternatively by using the diffraction method (Hilgart et al. 2011; Song et al. 2007). Crystal centering based on the diffracton method is especially attractive for micrometer-sized crystals. Optical centering of small crystals is challenging since visible light wavelengths (0.4–0.7 μm) are comparable to the crystal size and many crystals have irregular diffraction quality, which cannot be addressed by this technique. In diffraction-based crystal centering, the crystal is scanned in two dimensions using a small step size and at each step a diffraction image is taken, which is analyzed for locating and counting diffraction spots. The scored results are presented in a table which allows users to select optimally diffracting areas within the macroscopic sample (Cherezov et al. 2009).

Data collection and processing software packages

Typically, data extending to 2.5 Å resolution or higher are desirable for novel proteins and protein–ligand complexes, so that the model can be fitted unambiguously into the electron density map. However, in more challenging cases, data at 3 Å resolution or lower may be sufficient to fit the overall fold of a protein or the constituents of a multi-protein complex. A typical X-ray diffraction image, the electron density map to atomic resolution and the distribution of resolutions for protein structures in the PDB are depicted in Fig. 6. However, in many cases, diffraction properties of crystals are not known in advance, especially when crystals are small (in the micrometer range) and cannot be prescreened using in-house instrumentation prior to a synchrotron trip. It often takes a significant amount of time at the synchrotron to screen these sub-micron crystals

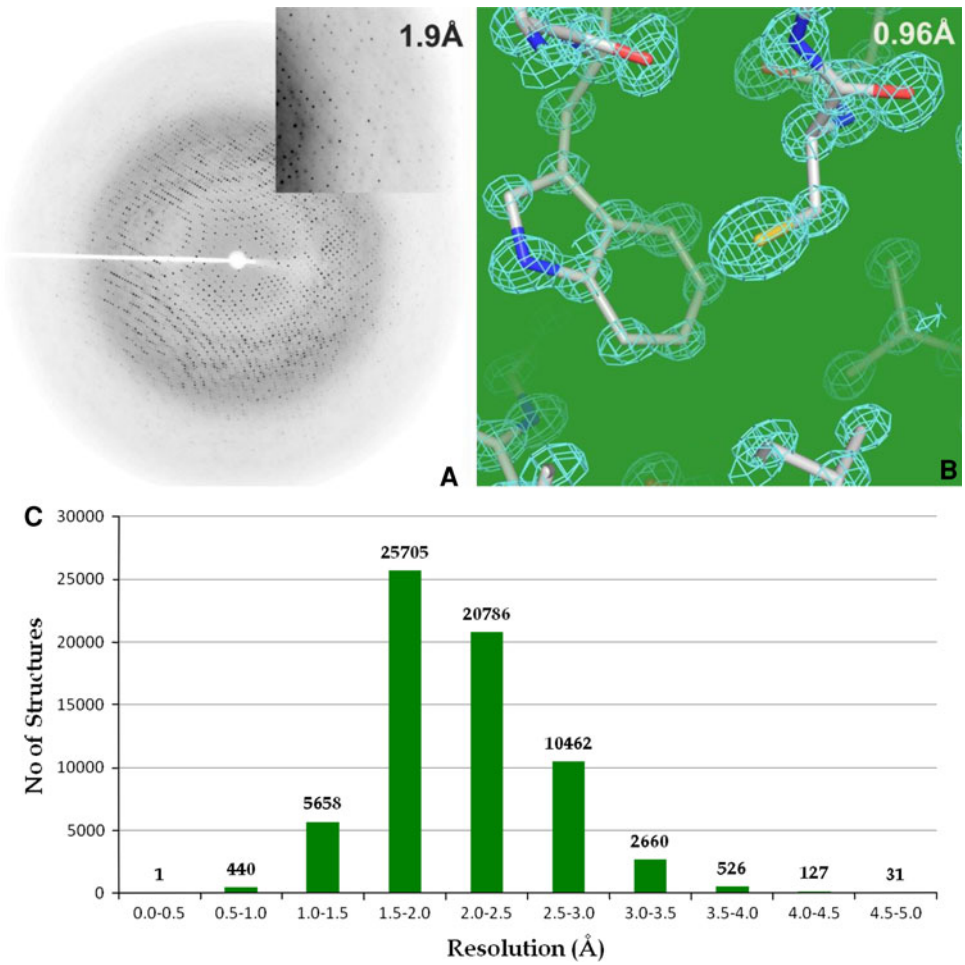

Fig. 6 Accuracy and details. **a** Representative X-ray diffraction pattern collected on a Marresearch GmbH imaging plate system. The diffraction extends to a maximum of 1.9 Å resolution at the edge of the image. **b** Representative portion of an electron density map at 0.96 Å resolution. The sticks represent the individual atoms for the amino acids that constitute the protein (carbon, *gray*; nitrogen, *blue*; oxygen, *red*; sulfur, *yellow*) and the chicken wire represents the corresponding experimental electron density for these atoms. **c** Histogram depicting the distribution of resolutions for protein structures in the PDB as of August 2011

to identify a well-diffracting crystal suitable for data collection. Whilst collecting data at the synchrotron beamline, the user must make decisions about the parameters of the experiment—exposure time, rotation range, oscillation angle, detector distance, beam size and wavelength—based on their experience, visual inspection of the diffraction images and information output by data-processing packages. Most of the instrumentation in the experimental station is computationally controlled using software packages such as *Blu-Ice* (McPhillips et al. 2002), *CBASS* (Skinner et al. 2006), *MxCube* (Gabadinho et al. 2010) and *JBlue-Ice* (Stepanov et al. 2011). However, very often an intuitive decision is made by the user on the exposure time to use. In cases where this has been overestimated, it can lead to significant radiation damage before the completion of data collection. In addition, an inappropriate data collection strategy can lead to the failure of an experiment. Computationally efficient modeling of the data statistics for any combination of data collection parameters provides a foundation for making a rational choice. The modeling of data statistics using a few test images allows one to quantitatively select which screened crystal gives the highest resolution using an appropriate rotation range and X-ray radiation dose prior to data collection (Bourenkov and Popov 2006, 2010).

The evaluation of the collected reflection intensities on the diffraction images involves the integration of the total intensity within all pixels of the individual spot profiles. The crystallographic program *HKL2000* is capable of carrying out data processing automatically (Borek et al. 2003; Minor et al. 2006). Other commonly used data-processing packages include *XDS* (Kabsch 1993) and *MOSFLM* (Leslie 2006). These programs all give excellent results with high-quality diffraction data, although their treatment of imperfect data differs owing to different approaches to indexing, spot integration and the treatment of errors. These programs can process data from a wide variety of modern area detectors from manufacturers including MarResearch, Rigaku/MSC, ADSC and MacScience. All these programs require crystallographers to make informative decisions and to input the correct experimental parameters to process the data successfully. There are ongoing activities at several synchrotron beamlines to develop expert systems that aim to automate the data collection strategy using the software *BEST* (Bourenkov and Popov 2006), *RADDOSE* (Paithankar and Garman 2010), *MOSFLM* and *XDS* to reduce the time required to successfully collect high-quality X-ray data.

Post-crystallization treatments to improve the quality of diffraction

Among the biggest problems in macromolecular crystallography is the relatively weak diffraction power of protein crystals and their sensitivity to ionizing radiation damage. Cryogenic methods provide great advantages in macromolecular crystallography, especially when synchrotron radiation is used for diffraction data collection. Apart from reducing the problem with radiation damage and enabling the storage and safe transport of frozen crystals, there are a number of additional benefits. For example, cryo-freezing can be exploited to trap normally unstable intermediates in enzyme-catalyzed reactions to permit their characterization. In addition, cryo-freezing can dramatically improve diffraction properties by reducing thermal vibrations and conformational disorder within the crystal, provided that the crystal is amenable to freezing and a suitable cryoprotectant has been selected. Of primary practical importance is the decrease in secondary radiation damage in the crystal caused by the diffusion of free radicals, typically permitting a complete data set to be collected from a single crystal. Cryogenic data collection has allowed efficient phasing using multi-wavelength methods.

When a crystal of a biological macromolecule is cooled to cryogenic temperatures, the main difficulty is to avoid the crystallization of any water present in the system, whether internal or external. Therefore, a cooling procedure has to be chosen that leads to a glass-like amorphous phase of the solvent. In principle, there are four options: (1) cooling on a timescale too fast for ice formation to occur (Hartmann et al. 1982), (2) cooling at high pressure by which the formation of the common hexagonal form of ice is circumvented (Thomanek et al. 1973), (3) replacing the liquid surrounding the crystal with a water-immiscible hydrocarbon oil such as *paratone-N* (Hope 1988, 1990), *paraffin oil* (Riboldi-Tunnicliffe and Hilgenfeld 1999) and *LV CryoOil*™ (MiTeGen), (4) modifying the physicochemical properties of the solvent by the addition of cryoprotectants in a way that a vitrified state can be reached at moderate cooling rates.

To prevent the nucleation of ice crystals, the last method is currently the most widely used. The crystal is permeated with a diffusible solvent containing cryoprotectants such as glycerol, sucrose or other organic solvents (Garman 1999; Garman and Owen 2005; Heras and Martin 2005). Determining the initial and optimal cryoprotectant concentration is often a process of trial and error. One must find suitable cryoprotectant concentrations that do not destroy the crystalline order while, at the same time, allowing the solvent to form an amorphous glass upon rapid cooling. Recently, trimethylamine *N*-oxide (TMAO) has been shown as a very versatile cryoprotectant for macromolecular crystals (Mueller-Dieckmann et al. 2011).

It has been shown that diffraction properties of flash-cooled macromolecular crystals can often be improved by warming and then cooling a second time—a procedure known as crystal annealing. Two different crystal-

annealing protocols have been reported (Garman 1999; Harp et al. 1998, 1999; Samygina et al. 2000; Yeh and Hol 1998) and many variants of these have been tried in the field. The first method involves removing a flash-cooled crystal from the cold gas stream and placing it in a cryo-protectant solution (either glycerol, MPD or Paratone-N oil) for several minutes before refreezing (Harp et al. 1998, 1999). There are several examples cited in the literature where this technique has been successfully applied (Felts et al. 2006; Manjasetty et al. 2001). In the second method, the cold stream is blocked for a fixed amount of time before the crystal is allowed to re-cool (Yeh and Hol 1998). Both annealing protocols can improve crystal resolution and mosaicity, although substantial crystal-to-crystal and mol-ecule-to-molecule variability has also been observed. Recently, the flash annealing technique has been automated using a cryo-shutter (Vahedi-Faridi et al. 2005), a device that blocks the 100 K nitrogen stream that bathes the crystal for a specific amount of time. The main advantage of the shutter system is that it allows a controlled, instant re-cooling of the crystal and the user can perform the flash annealing experiment remotely without entering the experimental hutch.

Diffraction quality can also be improved by post-crys-tallization treatments, such as controlled dehydration (Heras et al. 2003), to attempt to improve the crystal dif-fraction properties. A user-friendly apparatus for crystal dehydration has been designed and implemented at the ESRF/EMBL beamlines (Russi et al. 2011; Sanchez-Weatherby et al. 2009). In addition, Proteros biostructures GmbH has developed a *Free Mounting System* (*FMS*^{TM}) that precisely controls the humidity around a crystal, which can lead to dramatically improved diffraction data.

Remote data collection

Synchrotron data collection can be performed remotely from home institutions by accessing the instrumentation via advanced software tools that enable the network-based control of beamlines (Gonzalez et al. 2005). Remote access to synchrotron sources is becoming more popular, since it saves both time and resource and results in more efficient use of the beamtime. "Mail-in" crystallography (diffrac-tion data measured by synchrotron staff) is another popular option for X-ray diffraction data collection, whereby users ship their crystals to the synchrotron for data collection by the beamline scientists.

X-ray structure determination

Amplitudes or intensities can be measured directly from the X-ray diffraction experiment, but information relating to their relative phases cannot be measured. To be able to calculate an electron density map and subsequently deter-mine the protein structure, an estimate of the phases has to be obtained indirectly using mathematical approaches and this represents the *phase problem* in protein crystallography.

Structure determination methods

Heavy-atom incorporation (isomorphous replacement, anomalous scattering and anomalous dispersion), molecu-lar replacement and direct methods are commonly used techniques to solve protein structures. The general requirement for the exploitation of the anomalous signal for the determination of phase estimations via *multiple* or *single-wavelength anomalous diffraction* (*MAD* or *SAD*) techniques is that the protein crystal should contain anomalously scattering atoms, e.g., Hg, Pt or Se. With the advent of tunable X-ray sources and improved data col-lection techniques, it is now possible to measure the intensities of diffracted X-rays with very high precision. The small differences in intensities between Bijovoet pairs due to the presence of heavy atoms can be used to calculate initial estimates of the protein phase angle. One of the strategies widely used for the determination of novel pro-tein structures is selenomethionine incorporation, where selenomethionine is replaced by methionine in the protein during expression. This method has revolutionized protein X-ray crystallography and it is estimated that over two-thirds of all novel crystal structures have been determined using either Se-SAD or Se-MAD (Fig. 7a). Novel struc-tures can also be solved using the weak anomalous signals from atoms, such as sulfur and phosphorous present in certain macromolecules. SAD represents the most com-monly used technique for novel proteins in SP centers. *Multiple* or *single isomorphous replacement* (*MIR* or *SIR*) methods also require the introduction of heavy atoms such as mercury, platinum, uranium or gold into the macro-molecule under investigation. These heavy atoms must be incorporated into protein crystals without disrupting the lattice interactions so that it remains *isomorphous* with respect to the native crystal. In the *SIR* method, intensity differences between the heavy-atom derivatized and native crystal are used to calculate experimental phases. Recently, the *SIR* phasing protocol has been re-applied in the *radi-ation damage-induced phasing* (*RIP*) technique, where the differences in intensities induced by radiation damage are used as a phasing tool (Ravelli et al. 2003). Limitations of these phasing protocols are mainly due to the deleterious effect that a high X-ray dose has on a protein crystal. X-ray radiation damage induces many changes to the protein structure and to the solvent, resulting in a consistent number of damaged sites and a decrease in the diffraction quality of the crystal. As an alternative to X-rays, *ultravi-olet* (*UV*) radiation has been used to induce specific

changes in the macromolecule, which only marginally affects the quality of the diffraction (Nanao and Ravelli 2006) while inducing more selective changes to the protein structure. This method is known as *UV-RIP* (*ultraviolet radiation damage-induced phasing*). The most striking effect of UV radiation damage on protein crystals, as for X-ray radiation, is the breakage of disulfide bonds. Furthermore, this technique has been extended to a non-disulfide-containing protein, photoactive yellow protein, which contains a chromophore covalently attached through a thioester linkage to a cysteine residue (Nanao and Ravelli 2006) and to selenomethionine (MSe) proteins (Panjikar et al. 2011). Therefore, this method offers considerable potential, and selenium-specific UV damage could serve as an additional or even an alternative way of experimental phasing in macromolecular crystallography (de Sanctis et al. 2011). Another popular method adopted at SP centers is the use of iodide ion soaks and *SAD* experiments for de novo phasing (Abendroth et al. 2011).

Molecular replacement (*MR*) requires a search model for the protein under investigation, either determined from X-ray crystallography or from homology modeling, to calculate initial estimates of the phases of the new structure. The use of MR has become more commonplace with the expansion of the PDB and is currently used to solve up to 70% of deposited macromolecular structures where a homolog structure already exists (Fig. 7b: Pike et al. 2008). In cases when there are up to four molecules in the asymmetric unit of the crystal, the search model is structurally similar to the target protein and its oligomeric state is known, the MR method is fairly straightforward using programs such as *MOLREP* (Vagin and Teplyakov 1997), *AMoRe* (Navaza 2001) and *Phaser* (McCoy et al. 2007). To further streamline the MR procedure, a number of

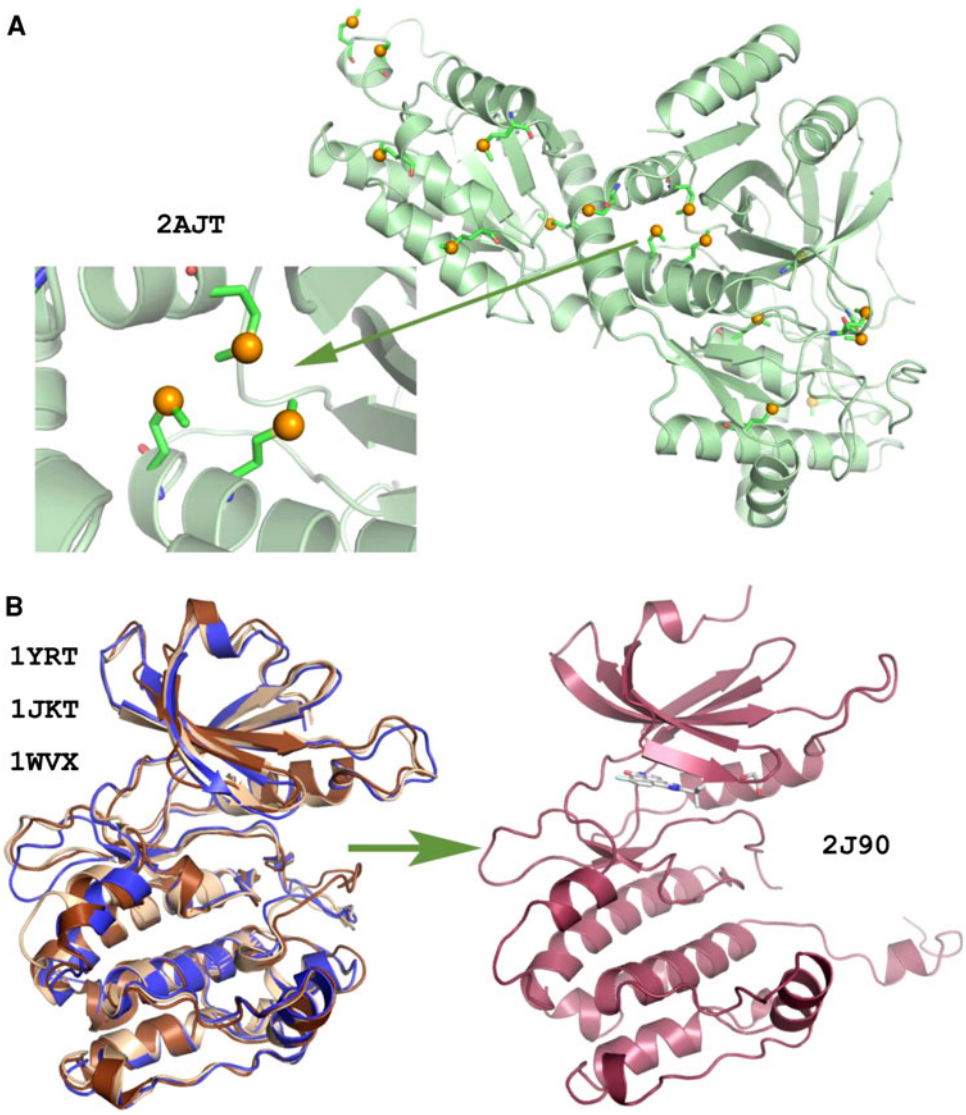

Fig. 7 Examples for widely used structure determination methods. **a** Structure of *E. coli* Arabinose Isomerase (PDB 2AJT) determined by *single-wavelength anomalous diffraction* (SAD). Selenomithionine residues are also shown (Manjasetty and Chance 2006). **b** Structure of DAPK3 (PDB 2J90) determined with the molecular replacement (MR) method using the template prepared by homolog structures (PDB 1YRT, 1JKT, 1WVX) (Pike et al. 2008)

automated MR pipelines have been developed. These include the Bias Removal Server (Reddy et al. 2003), *CaspR* (Claude et al. 2004), *BRUTEPTF* (Strokopytov et al. 2005) and MR pipeline (Schwarzenbacher et al. 2008). Other developments include *Auto-Rickshaw* (Panjikar et al. 2005) which is principally used for experimental phasing, but also uses phased MR as well as enabling a standard MR phasing protocol using *BALBES* (Long et al. 2008), *MrBUMP* (Keegan and Winn 2008) and a scheme for using comparative models in MR (Raimondo et al. 2007). Recently, MR phasing has been demonstrated for 2.0 Å data based on the combination of localizing model fragments such as small helices with *Phaser* and density modification with *SHELXE* (Rodriguez et al. 2009). In addition, improved MR by density-and energy-guided structure optimization has also been described (DiMaio et al. 2011).

It is worth noting that if an MR search is difficult primarily because the model is extremely poor and the resolution of the X-ray data is limited (lower than 2.0 Å), then the time spent attempting to obtain a solution with that model is usually inversely proportional to the usefulness of the solution once it has been obtained. This is partly because the model suffers from bias and often requires iterative, time-consuming manual correction using computer graphics in combination with model refinement. Interestingly, the determination of the substructure becomes easier when an anomalous difference Fourier synthesis can be calculated using preliminary phases from an MR solution. The subsequent use of this substructure to generate an unbiased electron density map (Baker et al. 1993) is often referred to as *MRSAD* (*molecular replacement with single-wavelength anomalous dispersion*) (Schuermann and Tanner 2003). A combination of MR and SAD has been automated and incorporated into the structure determination platform *Auto-Rickshaw*. The complete *MRSAD* procedure includes MR, model refinement, experimental phasing, phase improvement and automated model building; it has been shown that poor MR or SAD phases with phase errors larger than 70° can be improved using this described procedure (Panjikar et al. 2009) and a large fraction of the model can be determined in a purely automatic manner from X-ray data extending to better than 2.6 Å resolution.

Computational resources for structure determination

New software packages have been developed for determining the 3D structure of proteins to meet the HT requirements of SP projects (Jain and Lamour 2010). The software pipelines have varying degrees of automation deriving from different aims, but all require minimum user input to facilitate the automated location of heavy-atom sites, phase determination and phase improvement by solvent flipping/flattening, model building and refinement. Ideally, the structure solution process should be carried out in parallel with data processing. For instance, the most recent iteration of the *HKL* suite, *HKL-3000*, is capable of integrating automated data collection, processing, structure solution and refinement steps (Minor et al. 2006). The *Auto-Rickshaw* suite incorporates many widely used programs for automatic protein structure determination (Panjikar et al. 2005). *AutoSHARP* (Vonrhein et al. 2007), *CRANK* (Ness et al. 2004; Pannu et al. 2011), *BnP* (Weeks et al. 2002), *HKL2MAP* for *SHELX* (Pape and Schneider 2004; Schneider and Sheldrick 2002) and the *PHENIX* suite (Adams et al. 2004) are highly automated and provide all the tools necessary to proceed from substructure solution and phasing through to displaying and interpreting the resultant electron density map. In addition, these programs provide automated protocols to enable protein models to be built rapidly without user intervention, providing feedback on the success of the experiment while the crystal is still at or near the beamline. *AutoSHARP* includes various *CCP4* (Collaborative Computational Project, number 4) programs (Winn et al. 2011), uses the *SHELXD* software for locating heavy atoms and carries out density modification using either *DM* (Cowtan and Zhang 1999) or *SOLOMON* (Abrahams and Leslie 1996), while *ARP/wARP* (Langer et al. 2008; Morris et al. 2003) or *BUCCANEER* (Cowtan 2006) are used for automated model building. This pipeline can be run without user intervention once suitable input has been provided and can be rerun from any of the structure solution steps by the user whenever desired. The *CRANK* package invokes *BP3* (Pannu and Read 2004), *CRUNCH2* (de Graaff et al. 2001), *SHELXD*, *SOLOMON*, *DM*, *RESOLVE* (Terwilliger 2000), *BUCANEER* and *ARP/wARP* along with a few *CCP4* programs and uses standard XML input at every step of the structure solution. The process may be invoked either using the *CCP4* graphical user interface (*CCP4i*) or off-line and the user must choose the defined path through the pipeline. The *BnP* pipeline includes *SnB* (Xu and Weeks 2008) and the *PHASE* package for structure solution. *HKL-3000* is a commercially available software package, which includes the data-processing programs *DENZO/SCALEPACK* (Otwinowski and Minor 1997) along with structure solution programs including modified versions of *MLPHARE*, *SHELXC/D/E*, *DM* and *ARP/wARP*. The *PHENIX* software suite is a highly automated system for macromolecular structure determination that can rapidly arrive at an initial partial model of a structure without significant human intervention, given moderate resolution and good quality data. The *Auto-Rickshaw* pipeline has been developed with its primary aim to validate the X-ray diffraction experiment while the crystal is still at or near the synchrotron

beamline. The software pipeline is optimized for speed so that the user has the ability to evaluate their data in the minimum possible time. *Auto-Rickshaw* makes use of publicly available macromolecular crystallography software. The entire process in the pipeline is fully automatic. Each step of the structure solution process is governed by the decision-making module within *Auto-Rickshaw*, which attempts to mimic the decisions of an experienced crystallographer for a number of phasing protocols (SAD, MAD, RIP, MR and variations thereof). Once the input parameters (number of amino acids, heavy atoms, molecules per asymmetric unit, probable space group and phasing protocol) and X-ray intensity data have been input into *Auto-Rickshaw*, no further user intervention is required. It proceeds step by step through the structure solution using the decision makers. In cases where a problem is encountered during the structure solution process, the user is informed so that the data collection, the data quality, space group ambiguity or optimization of the anomalous signal is flagged as a problem. Once all the steps have been run successfully, *Auto-Rickshaw* provides a tarball, which includes all the necessary files to evaluate the electron density map and model, including ready-made scripts for the graphics programs *COOT* (Emsley and Cowtan 2004; Emsley et al. 2010), *O* (Jones et al. 1991) and *XtalView* (McRee and Israel 2008). The *Auto-Rickshaw* server (http://www.embl-hamburg.de/Auto-Rickshaw) is freely accessible to the SP community to aid in their protein structure determination effort.

Structure and function

Proteins play key roles in almost every biological process and participate in a variety of physiological functions. The key to unraveling how proteins perform their different roles lies in understanding the relationship between protein structure and function.

Computational tools and Web servers

A wiki, *Proteopedia*, provides the forum to contribute and share information about a particular 3D structure to exploit its functional role through collaboration, follow-up studies and joint publication (Hodis et al. 2010; Prilusky et al. 2011). Many of the protein structures determined at SP centers are 'hypothetical proteins' of unknown function. Insights into the biochemical function of these proteins can be derived by bioinformatics tools such as *TOPSAN* (*The Open Protein Structure Annotation Network*), a Web-based platform which facilitates collaborations that have resulted in insightful structure–function analysis for many proteins leading to numerous peer-reviewed publications (Ellrott et al. 2011; Weekes et al. 2010). A number of other servers, such as *ProFunc* (Laskowski et al. 2005), *3D-Fun* (von Grotthuss et al. 2008) and *ProTarget* (Sasson and Linial 2005), enable automated protein function annotations using protein structures. These servers accept protein coordinates in the standard PDB format and compare them with all known protein structures in the PDB. If structural hits are found for proteins of known function, they are listed together with their function and some vital comparison statistics. A recently developed platform called iSee (*interactive Structurally enhanced experience*) allows the interaction and annotation of novel structures and provides a powerful tool for disseminating the full range of structural information. Interestingly, this platform is hosting and featuring the animations adopted by journals that 'fly' the reader through structural representations to specific molecular features (Davis et al. 2010; Rellos et al. 2010). This novelty in exposing structural knowledge across the life sciences is one of the breakthroughs attributable to SP.

Protein fold, sequence motif, binding site, oligomeric state and surface features for protein annotation

The 3D folding patterns of proteins can be categorized into classes which allow inferences to be made about their molecular function. A comparison of the 3D structures of proteins of known function in the PDB has shown that proteins sharing common folds are often evolutionarily related. *Protein folds* are more highly conserved over time than protein sequence, and structural similarities between a protein of known function and a novel (hypothetical) protein implies that these proteins have related function (Chothia and Lesk 1986). The most commonly used servers for comparing the fold of a newly determined structure against structures in the PDB are *DALI* at EMBL (Holm and Sander 1999) and *VAST* (*Vector Alignment Search Tool*) at NCBI (Gibrat et al. 1996). For example, the hypothetical protein FLJ36880 has a fold which indicates that it is a member of the fumarylacetoacetate hydrolase family (Manjasetty et al. 2004b) and the fold of MJ0882 highly resembles that of a methyltransferase despite limited sequence similarity to any known methyltransferase (Huang et al. 2002). *Structural motifs* in a protein usually represent regions characteristic of specific functions (Godsey et al. 2007). For example, helix–turn–helix motifs are highly conserved in DNA binding proteins. Similar functional sites can be found across different folds in proteins as a result of convergent evolution. This is frequently observed across folds involving *metal ions*. The server *PINTS* (*Patterns in Non-homologous Tertiary Structures*) helps to identify functional patterns in non-homologous structures (Stark and Russell 2003). For example, the crystal structure of YodA from *E. coli* indicates that it is a metal-binding lipocalin-like protein and it may be important in the metal stress

response (David et al. 2003). In addition, the hypothetical protein ybeY belongs to the UPF0054 family and binds a metal ion. Its structure and sequence similarity to a number of predicted metal-dependent hydrolases provides a functional assignment for this protein (Zhan et al. 2005). The crystal structure of *Homo sapiens* PTD012 reveals a zinc-containing hydrolase fold (Manjasetty et al. 2006).

Proteins are a unique class of biological molecules in that they can recognize and interact with diverse substances. They contain complementary clefts and surfaces designed to bind to specific molecules. The identification of binding sites on the surface of proteins can give insights into biological function. There are many programs available for detecting and visualizing binding sites on the surface of proteins including *SURFNET* (Laskowski 1995), *CAVER* (Petrek et al. 2006) and *dxTuber* (Raunest and Kandt 2011). For example, the crystal structure of TTHB192 does not have the same signature sequence motif as the RNA recognition motif domain, however, the presence of an evolutionarily conserved basic patch on the β-sheet could be functionally relevant for nucleic acid binding (Ebihara et al. 2006). The structure for a representative of UPF0044, *E. coli* YhbY, possesses an IF3C-like core fold which is common in several RNA-binding proteins. Members of UPF0044 possess a basic surface on their β-sheet face and a proximal GKxG loop that are suggestive of an RNA recognition surface (Ostheimer et al. 2002). If the *binding site is occupied by a small molecule* carried over from the crystallization buffer or protein purification, then it is often easy to detect the functional "hot spot" in the protein structure. For example, TM841 is a 35-kDa protein comprising two separate domains with a novel fold. Therefore, it was not possible to derive any clues about this protein's function from a comparison with proteins in the PDB. However, the electron density map clearly showed the presence of a fatty acid molecule bound in a pocket between the two protein domains, suggesting that TM841 may play a role in fatty acid transport or metabolism (Schulze-Gahmen et al. 2003). In a second example, ligands bound to the structures of p14.5, TdcF and Rv2704, which belong to the highly conserved YjgF/YER057c/UK114 protein superfamily, clearly indicated the presence of a functionally important substrate binding site. The structure of human p14.5 contains at least one benzoic acid molecule per site forming bi-dentate interactions between its carboxylate moiety and the guanidinium group of a strictly conserved arginine, Arg107 (Manjasetty et al. 2004a). Furthermore, the bonds to the ligands in the TdcF crystal structure clearly highlight the importance of the conserved residues (Burman et al. 2007). Residue conservation in an amino acid sequence or related proteins is a strong indicator of functionally important sites. Given a multiple sequence alignment of one protein against all related proteins, it is possible to determine the level of sequence conservation. Once a 'conservation score' for each residue in the protein has been calculated, it is useful to map these scores onto its 3D structure to see whether certain regions of the structure are more highly conserved than others. It has been demonstrated that clusters of highly conserved residues on the surface of proteins can reliably identify ligand-binding sites or protein–protein interaction sites. For example, the amino acid conservation pattern and observed metal-binding site in the crystal structure of *E. coli* YcDX indicates the location of its putative active site. All residues involved in metal coordination are invariant in the YcdX family, highlighting their functional importance (Teplyakov et al. 2003). An in-depth study is warranted to check the reliability of these ligand-binding discovery programs. The ligand or metal-binding sites observed through computational approaches can be further verified by obtaining experimental data. For example, in a recent study, a combination of experimental and bioinformatics approaches has provided a comprehensive active site analysis on the genome scale for metalloproteins, revealing new insights into their structure and function (Shi et al. 2011).

The arrangement of molecules in the asymmetric unit of the crystal and analysis of the accessible surface area buried at the subunit interfaces can be used to identify the most likely biological unit (*oligomeric state*) for a protein. The evolutionarily conserved trimeric structure of CutA1 proteins suggests a role in signal transduction (Arnesano et al. 2003). The functional annotation of the large number of structures of hypothetical proteins was published based on a bioinformatics approach (Shin et al. 2007; Yakunin et al. 2004). Therefore, crystal structures can provide insights into biological function even in the absence of any other biochemical data. However, the protein structure alone will provide conclusive functional annotation only in a limited number of cases. Therefore, SP centers collaborate with academic laboratories to resolve important questions in biology and disease. In this approach, highly organized networks of investigators apply the new paradigm of HT structure determination, which has been successfully developed at SP centers during the past decade, to study a broad range of important biological and biomedical problems. One such example is the NIH-funded *Enzyme Function Initiative* (http://enzymefunction.org).

Information management

SP data are hard to manage due to their complexity and the fact that different parts of the structure determination process on the same protein target are often performed at different laboratories as a collaborative effort. The file-based communications and data transactions become time consuming and sometimes unmanageable. Efficient data

management techniques have been developed at SP centers to keep track of the enormous amount of data generated, which minimize duplication of effort and maximize the chances of success at each step (Haquin et al. 2008). Notable laboratory information management systems (LIMS) of SP centers include *Sesame* (Zolnai et al. 2003), *HalX* (Prilusky et al. 2005), *PiMS* (Morris et al. 2011), *SPEX Db* (Raymond et al. 2004) and *IceDB*. All the SP centers have developed their own data management systems and are linked to a centralized target registration database, *TargetDB* (Chen et al. 2004) hosted by the PDB. *PepcDB*, an extension of *TargetDB*, allows groups to report protocols and experimental details (Pan et al. 2007). Subsequently, the *Protein Structure Initiative Knowledgebase* (PSIKB) was created by integrating *TargetDB* and *PepcDB* to turn the products of the SP effort into knowledge that can be accessed by the life sciences community (Berman et al. 2009). Furthermore, a *Structural Biology Knowledgebase* (SBKB) web resource (http://www.sbkb.org) has recently been developed to aid protein research with improved features to foster collaborations between the biological community and SP centers (Gabanyi et al. 2011).

Structural proteomics for biology

Structural knowledge of a protein clearly provides clues relating to its biological activity and physiological role. SP is one of the recent technologies that promotes drug discovery and biotechnological applications. Structural information can be used in many ways to ascertain the functional properties of cellular components. One of the crucial components for understanding the functions of novel proteins is the analysis of their experimental or modeled 3D structures. SP centers have provided an enormous impetus for methods development in structural biology and many laboratories are now actively implementing these technologies.

Follow-on research

Scientists and engineers are now involved in utilizing structural knowledge of proteins generated by the SP approach as a basis for understanding protein function to utilize proteins in various technological applications as follow-on research. For instance, in Europe, the emphasis of the *Structural Proteomics IN Europe* (SPINE2; http://www.spine2.eu/SPINE2/index.jsp) initiative has been to apply HT technologies to systems of biological interest, the ultimate aim being to solve significant problems more effectively. Recently, *INSTRUCT* (http://www.structuralbiology.eu/) has started offering scientists access to world-class structural biology and SP infrastructures and expertise that makes such integration possible more rapidly, creating a coherent forum for structural biology. This forum will stimulate closer collaboration between scientific communities and initiatives in the life sciences. For instance, in Germany, one of the *INSTRUCT* centers, the *Protein Sample Production Facility* (http://www.pspf.de) of the Helmholtz association, provides HT *E. coli* protein expression and large-scale production technology with animal cell lines for European structural biologists. The *HT Crystallization platform* at the EMBL Grenoble Outstation, France, offers automated crystallization to European researchers. Eleven facilities from across Europe provide installations for applications including *macromolecular X-ray crystallography data collection*. *BioStruct-X* (http://www.biostructx.org/) cooperates with *INSTRUCT* aiming to provide an integrated and coordinated technology platform to all relevant methods in structural biology.

In the USA, the goal of the PSI:Biology project is to apply the paradigm of HT structure determination via highly organized networks of investigators to solve the 3D structure of proteins and macromolecular complexes representing significant biological and biomedical problems. For instance, a protein family specific platform, *GPCR* (*G-protein coupled receptor*) network (http://gpcr.scripps.edu/) was established to determine the high-resolution structure and function of GPCRs distributed broadly across the phylogenetic family tree. GPCRs mediate many important cellular signal transduction events related to differentiation, proliferation, angiogenesis, cancer, development and cell survival. GPCRs represent the targets for 60–70% of drugs currently in development. The recent progress and success in the structure determination of GPCRs by utilizing SP technologies like LCP (Lipidic Crystal Phase) crystallization (Cherezov 2011) and microcrystallography technologies for data collection (10 × 10 μm beam size, microfocus beamline, Pilatus 6 M detector) have been described (Cherezov et al. 2009; Shimamura et al. 2011). In addition, HT-enabled structural biology partnerships have also been established. For example, the *IFN* (Immune Function Network), a consortium of immunologists, geneticists, computational biochemists and HT structural biologists, is committed to the coordinated structural, in vitro biochemical and in vivo functional analyses of secreted molecules and ectodomains of cell surface molecules which control adaptive and innate immunity (Chattopadhyay et al. 2009). The research activities of the IFN will be conducted in collaboration with the NYSGRC (http://www.nysgxrc.org/). The *TB Structural Genomics Consortium* is a worldwide consortium of scientists developing a foundation for tuberculosis diagnosis and treatment by determining the 3D structures of proteins from *Mycobacterium tuberculosis* (Mtb). The

consortium seeks to solve structures of proteins that are of great interest to the TB biology community (http://www.webtb.org/) (Musa et al. 2009). Tuberculosis poses a global health emergency, which has been compounded by the emergence of drug-resistant Mtb strains. For instance, the protein structure of isocitrate lyase, a persistence factor of Mtb (Sharma et al. 2000), has been extensively studied and advances in technology have enabled the assembly of HT pipelines that can be used for the development of glyoxylate cycle inhibitors as new drugs for the treatment of this disease (Munoz-Elias and McKinney 2005). Further, as a step toward the better integration of protein 3D structural information in cancer systems biology, NESG has constructed a *HCPIN* (*Human Cancer Pathway Protein Interaction Network*) by analzsing several classical cancer-associated signaling pathways and their physical protein–protein interactions (Huang et al. 2008).

Systems biology and biotechnology

In systems biology, proteins are visualized as a network of interconnected dots. To understand the complexity of cellular function, one should know the detailed 3D behaviors of all the available dots which form the basis of life. Furthermore, the structures of these proteins could provide quantitative parameters to help elucidate functional networks through knowledge of protein function, evolution and interactions. The protein structures generated by SP can be used for the assignment of domain structure, functional annotation and the prediction of interaction partners in biochemical pathways (Harrill and Rusyn 2008). The structural information can be used to further characterize large-scale protein interaction networks by providing the key functional properties of cellular components (Beltrao et al. 2007).

Biotechnology embraces the bioproduction of fuels and chemicals from renewable sources. Sustainable energy is a major problem in the twenty-first century. If biofuels are to be part of the solution, this field must accept a degree of scrutiny unprecedented in the development of a new industry. That is because sustainability deals explicitly with the role of biofuels in insuring the well-being of our planet, our economy and our society, both today and in the future. The development of detailed kinetic models that include accurate regulatory network parameters will facilitate the identification of enzymatic bottlenecks in the metabolic pathways that could be harnessed to achieve biofuels overproduction. The latest advances in SP will continue to identify the biocatalysts, which power the development of enzyme reactors for producing substantial amounts of biofuels (Daniels et al. 2011). Some biomolecules are robust enough to be used in biotechnological applications. For instance, enzymes can be used to break down starch to form sweeteners. The structure–function relationship of *E. coli* arabinose isomerase, ECAI, advanced its application in tagatose (a new sweetener) production (Manjasetty et al. 2010; Manjasetty and Chance 2006) (Fig. 7a).

Nanobiotechnology is a novel branch of futuristic science and engineering. A nanobiomachine is a machine formed by a biomolecule with a nanoscale diameter. The knowledge of a protein sequence provides the basis for understanding these nanobiomachines, which ultimately describe its functional significance. The structural and functional knowledge of a protein is essential to utilize proteins in nanotechnology applications and to develop bionanodevices. Glucose oxidase is a small, stable enzyme that oxidizes glucose into glucolactone, converting oxygen into hydrogen peroxide in the process. It is used as the heart of *biosensors* that measure the amount of glucose in the blood. Insights from protein structure can be crucial in engineering proteins for nanotechnology applications.

Models of protein structures and drug design

SP projects around the globe were established to determine the structures of proteins in an HT, automated fashion. However, despite the advances made by SP organizations in terms of automation, throughput and methodology development, the structures of certain classes of proteins, such as membrane proteins, are still notoriously difficult to determine. This warrants alternative techniques to generate models for these proteins to enhance our understanding of their physical and chemical properties. This has led to the development of a large number of bioinformatics tools capable of generating models for these novel proteins. Among them, *MODELLER* (Eswar et al. 2008) and *ROSETTA* (Das and Baker 2008) represent two of the best protein structure prediction servers. *MODELLER* generates a model of an unknown protein using a template structure generated by the SP approach, whereas *ROSETTA* provides ab initio structure prediction of the unknown protein. Biomodeling provides the ability to understand the physicochemical properties of proteins of biomedical importance with undetermined 3D structures. It involves a range of computerized techniques based on quantum physics and experimental proteomics data to predict and correlate biological properties at the molecular level. Statistical and regression analysis techniques are the best methodologies and are capable of predicting geometries, energies, and electronic and spectroscopic properties. Homology-based modeling, as applied by *MODELLER* and *ROSETTA*, relies on sequence alignments between proteins of known structure and the target protein. The accuracy of the calculated model is dependent on the accuracy of the sequence alignment and the divergence between target and template. The most accurate alignments are obtained by iterative and

profile or *HMM* (*Hidden Markov Models*)-based methods. In addition, structural data can be used to verify and improve alignments.

Biomodeling has become a valuable and essential tool in the drug design and discovery process. Drug design is a 3D puzzle where small drug molecules are fitted into the active site of a protein. The factors which affect protein–ligand interactions can be characterized by molecular docking and studying quantitative structure activity relationships (*QSAR*). Traditionally, drug discovery relies on a stepwise synthesis and screening of large numbers of compounds to optimize drug activity profiles. The design of new and more potent drugs against diseases such as cancer, AIDS and arthritis can be aided using bioinformatics tools such as computer-assisted drug design (*CADD*) or computer-assisted molecular design (*CAMD*). Structural bioinformatics tools not only have the potential to build predictive models of the proteins of biomedical interest, but also help to bring new drugs to market. Complementary *in silico* methods, such as structure-based drug design (*SBDD*), incorporate the knowledge from high-resolution 3D protein structures generated by SP to probe structure–function relationships, identify and select therapeutically relevant targets (assessing druggability), study the molecular basis of protein–ligand interactions, characterize binding pockets, develop target-focused compound libraries, identify hits by HT docking (*HTD*) and optimize lead compounds, all of which can be used to rationalize and increase the speed and cost-effectiveness of the drug discovery process. An analysis of the results obtained by several docking and modeling programs has shown that, in most cases, they can work well. Most of the programs used in drug discovery have incorporated subroutines to identify false positives or negatives using scoring functions, which has led to a significant improvement in hit rates.

Conclusion

Contributions from SP are mainly twofold: first, novel structural information has been generated to understand the proteome of various organisms; second, innovative HT technologies have been developed for protein structure determination. These technologies and related structural information have, in turn, been exploited by biologists in many different ways in broad areas of life sciences research. However, structural knowledge alone is not sufficient to fully understand a protein's cellular role. Hence, a major bottleneck is studying a protein's behavior and dynamics within larger macromolecular assemblies and protein–protein interactions within a cellular pathway. Research such as this will drive the application of SP in the decades that follow.

Acknowledgments We thank Dr. J. Michael Sauder (Lilly Biotechnology Center, 10300 Campus Point Dr, San Diego, CA, USA) for critical reading and comments on this manuscript.

References

Abendroth J, Gardberg AS, Robinson JI et al (2011) SAD phasing using iodide ions in a high-throughput structural genomics environment. J Struct Funct Genomics 12:83–95. doi:10.1007/s10969-011-9101-7

Abrahams JP, Leslie AG (1996) Methods used in the structure determination of bovine mitochondrial F1 ATPase. Acta Crystallogr D Biol Crystallogr 52:30–42. doi:10.1107/S0907444995008754

Adams PD, Gopal K, Grosse-Kunstleve RW et al (2004) Recent developments in the PHENIX software for automated crystallographic structure determination. J Synchrotron Radiat 11:53–55

An Y, Meresse P, Mas PJ, Hart DJ (2011) CoESPRIT: a library-based construct screening method for identification and expression of soluble protein complexes. PLoS ONE 6:e16261. doi:10.1371/journal.pone.0016261

Aricescu AR, Assenberg R, Bill RM et al (2006) Eukaryotic expression: developments for structural proteomics. Acta Crystallographica Section D 62:1114–1124. doi:10.1107/S0907444906029805

Arnesano F, Banci L, Benvenuti M et al (2003) The evolutionarily conserved trimeric structure of CutA1 proteins suggests a role in signal transduction. J Biol Chem 278:45999–46006. doi:10.1074/jbc.M304398200

Baker D, Krukowski AE, Agard DA (1993) Uniqueness and the ab initio phase problem in macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 49:186–192. doi:10.1107/S0907444992008801

Beltrao P, Kiel C, Serrano L (2007) Structures in systems biology. Curr Opin Struct Biol 17:378–384. doi:10.1016/j.sbi.2007.05.005

Berger I, Fitzgerald DJ, Richmond TJ (2004) Baculovirus expression system for heterologous multiprotein complexes. Nat Biotechnol 22:1583–1587

Berman HM, Westbrook JD, Gabanyi MJ et al (2009) The protein structure initiative structural genomics knowledgebase. Nucleic Acids Res 37:D365–D368. doi:10.1093/nar/gkn790

Berrow NS, Büssow K, Coutard B et al (2006) Recombinant protein expression and solubility screening: a comparative study. Acta Crystallographica Section D 62:1218–1226

Beteva A, Cipriani F, Cusack S et al (2006) High-throughput sample handling and data collection at synchrotrons: embedding the ESRF into the high-throughput gene-to-structure pipeline. Acta Crystallogr D Biol Crystallogr 62:1162–1169. doi:10.1107/S0907444906032859

Bieniossek C, Nie Y, Frey D et al (2009) Automated unrestricted multigene recombineering for multiprotein complex production. Nat Meth 6:447–450. http://www.nature.com/nmeth/journal/v6/n6/suppinfo/nmeth.1326_S1.html

Borek D, Minor W, Otwinowski Z (2003) Measurement errors and their consequences in protein crystallography. Acta Crystallogr D Biol Crystallogr 59:2031–2038

Bourenkov GP, Popov AN (2006) A quantitative approach to data-collection strategies. Acta Crystallogr D Biol Crystallogr 62:58–64

Bourenkov GP, Popov AN (2010) Optimization of data collection taking radiation damage into account. Acta Crystallogr D Biol Crystallogr 66:409–419. doi:10.1107/S0907444909054961

Burley SK (2000) An overview of structural genomics. Nat Struct Biol 7(Suppl):932–934

Burman JD, Stevenson CE, Sawers RG, Lawson DM (2007) The crystal structure of TdcF, a member of the highly conserved YjgF/YER057c/UK114 family. BMC Struct Biol 7:30. doi:10.1186/1472-6807-7-30

Busso D, Peleg Y, Heidebrecht T et al (2011) Expression of protein complexes using multiple *Escherichia coli* protein co-expression systems: a benchmarking study. J Struct Biol 175:159–170. doi:10.1016/j.jsb.2011.03.004

Büssow K, Scheich C, Sievert V et al (2005) Structural genomics of human proteins–target selection and generation of a public catalogue of expression clones. Microb Cell Fact 4:21. doi:10.1186/1475-2859-4-21

Chance MR, Fiser A, Sali A et al (2004) High-throughput computational and experimental techniques in structural genomics. Genome Res 14:2145–2154

Chang VT, Crispin M, Aricescu AR et al (2007) Glycoprotein structural genomics: solving the glycosylation problem. Structure 15:267–273

Chattopadhyay K, Lazar-Molnar E, Yan Q et al (2009) Sequence, structure, function, immunity: structural genomics of costimulation. Immunol Rev 229:356–386. doi:10.1111/j.1600-065X.2009.00778.x

Chayen NE (1997) The role of oil in macromolecular crystallization. Structure 5:1269–1274

Chen L, Oughtred R, Berman HM, Westbrook J (2004) TargetDB: a target registration database for structural genomics projects. Bioinformatics 20:2860–2862. doi:10.1093/bioinformatics/bth300

Cherezov V (2011) Lipidic cubic phase technologies for membrane protein structural studies. Curr Opin Struct Biol 21:559–566. doi:10.1016/j.sbi.2011.06.007

Cherezov V, Hanson MA, Griffith MT et al (2009) Rastering strategy for screening and centring of microcrystal samples of human membrane proteins with a sub-10 microm size X-ray synchrotron beam. J R Soc Interface 6(Suppl 5):S587–S597. doi:10.1098/rsif.2009.0142.focus

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826

Cipriani F, Felisaz F, Launer L et al (2006) Automation of sample mounting for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 62:1251–1259. doi:10.1107/S0907444906030587

Claude JB, Suhre K, Notredame C et al (2004) CaspR: a web server for automated molecular replacement using homology modelling. Nucleic Acids Res 32:W606–W609. doi:10.1093/nar/gkh400.32/suppl_2/W606

Collinet B, Friberg A, Brooks MA et al (2011) Strategies for the structural analysis of multi-protein complexes: Lessons from the 3D-Repertoire project. J Struct Biol 175:147–158. doi:10.1016/j.jsb.2011.03.018

Cornvik T, Dahlroth SL, Magnusdottir A et al (2005) Colony filtration blot: a new screening method for soluble protein expression in *Escherichia coli*. Nat Methods 2:507–509

Coulibaly F, Chiu E, Ikeda K et al (2007) The molecular organization of cypovirus polyhedra. Nature 446:97–101. doi:10.1038/nature05628

Cowieson NP, Wensley B, Listwan P et al (2006) An automatable screen for the rapid identification of proteins amenable to refolding. Proteomics 6:1750–1757

Cowtan K (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr D Biol Crystallogr 62:1002–1011. doi:10.1107/S0907444906022116

Cowtan KD, Zhang KY (1999) Density modification for macromolecular phase improvement. Prog Biophys Mol Biol 72:245–270 pii:S0079-6107(99)00008-5

D'Arcy A, Villard F, Marsh M (2007) An automated microseed matrix-screening method for protein crystallization. Acta Crystallogr D Biol Crystallogr 63:550–554. doi:10.1107/S0907444907007652

Daniels C, Michan C, Ramos JL (2011) Microbial Biotechnology: biofuels, genotoxicity reporters and robust agro-ecosystems. Microb Biotechnol 3:239–241. doi:10.1111/j.1751-7915.2010.00177.x

Das R, Baker D (2008) Macromolecular modeling with rosetta. Annu Rev Biochem 77:363–382. doi:10.1146/annurev.biochem.77.062906.171838

David G, Blondeau K, Schiltz M et al (2003) YodA from *Escherichia coli* is a metal-binding, lipocalin-like protein. J Biol Chem 278:43728–43735. doi:10.1074/jbc.M304484200

Davis SJ, Puklavec MJ, Ashford DA et al (1993) Expression of soluble recombinant glycoproteins with predefined glycosylation: application to the crystallization of the T-cell glycoprotein CD2. Protein Eng 6:229–232

Davis TL, Walker JR, Campagna-Slater V et al (2010) Structural and biochemical characterization of the human cyclophilin family of peptidyl-prolyl isomerases. PLoS Biol 8:e1000439. doi:10.1371/journal.pbio.1000439

de Graaff RA, Hilge M, van der Plas JL, Abrahams JP (2001) Matrix methods for solving protein substructures of chlorine and sulfur from anomalous data. Acta Crystallogr D Biol Crystallogr 57:1857–1862 pii:S0907444901016535

de Sanctis D, Tucker PA, Panjikar S (2011) Additional phase information from UV damage of selenomethionine labelled proteins. J Synchrotron Radiat 18:374–380. doi:10.1107/S0909049511004092

Derewenda ZS (2010) Application of protein engineering to enhance crystallizability and improve crystal properties. Acta Crystallogr D Biol Crystallogr 66:604–615. doi:10.1107/S090744491000644X

Dessau MA, Modis Y (2011) Protein crystallization for X-ray crystallography. J Vis Exp. doi: 10.3791/2285

DiMaio F, Terwilliger TC, Read RJ et al (2011) Improved molecular replacement by density- and energy-guided protein structure optimization. Nature 473:540–543. doi:10.1038/nature09964

Dyson MR (2010) Selection of soluble protein expression constructs: the experimental determination of protein domain boundaries. Biochem Soc Trans 38:908–913. doi:10.1042/BST0380908

Ebihara A, Yao M, Masui R et al (2006) Crystal structure of hypothetical protein TTHB192 from *Thermus thermophilus* HB8 reveals a new protein family with an RNA recognition motif-like domain. Protein Sci 15:1494–1499. doi:10.1110/ps.062131106

Echalier A, Glazer RL, Fulop V, Geday MA (2004) Assessing crystallization droplets using birefringence. Acta Crystallogr D Biol Crystallogr 60:696–702. doi:10.1107/S0907444904003154

Ellrott K, Zmasek CM, Weekes D et al (2011) TOPSAN: a dynamic web database for structural genomics. Nucleic Acids Res 39:D494–D496. doi:10.1093/nar/gkq902

Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 60:2126–2132

Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Crystallographica Section D Biological Crystallography 66:486–501. doi:10.1107/S0907444910007493

Eswar N, Eramian D, Webb B et al (2008) Protein structure modeling with MODELLER. Methods Mol Biol 426:145–159. doi:10.1007/978-1-60327-058-8_8

Felts RL, Reilly TJ, Calcutt MJ, Tanner JJ (2006) Cloning, purification and crystallization of *Bacillus anthracis* class C

acid phosphatase. Acta Crystallogr Sect F Struct Biol Cryst Commun 62:705–708. doi:10.1107/S174430910602389X
Forsythe E, Achari A, Pusey ML (2006) Trace fluorescent labeling for high-throughput crystallography. Acta Crystallogr D Biol Crystallogr 62:339–346. doi:10.1107/S0907444906000813
Gabadinho J, Beteva A, Guijarro M et al (2010) MxCuBE: a synchrotron beamline control environment customized for macromolecular crystallography experiments. J Synchrotron Radiat 17:700–707. doi:10.1107/S0909049510020005
Gabanyi MJ, Adams PD, Arnold K et al (2011) The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. J Struct Funct Genomics 12:45–54. doi:10.1007/s10969-011-9106-2
Garman E (1999) Cool data: quantity and quality. Acta Crystallogr D Biol Crystallogr 55:1641–1653 (BA0025)
Garman EF, Owen RL (2005) Cryocooling and radiation damage in macromolecular crystallography. Acta Crystallographica Section D Biological Crystallography 62:32–47. doi:10.1107/S0907444905034207
Gerdts CJ, Elliott M, Lovell S et al (2008) The plug-based nanovolume microcapillary protein crystallization system (MPCS). Acta Crystallogr D Biol Crystallogr 64:1116–1122. doi:10.1107/S0907444908028060
Gibrat JF, Madej T, Bryant SH (1996) Surprising similarities in structure comparison. Curr Opin Struct Biol 6:377–385
Godsey MH, Minasov G, Shuvalova L et al (2007) The 2.2 A resolution crystal structure of *Bacillus cereus* Nif3-family protein YqfO reveals a conserved dimetal-binding motif and a regulatory domain. Protein Sci 16:1285–1293. doi:10.1110/ps.062674007
Gonzalez A, Cohen A, Eriksson T et al (2005) Remote Access to the SSRL macromolecular crystallography beamlines. Syncrotron Radiation News 18:36–39
Groves MR, Muller IB, Kreplin X, Muller-Dieckmann J (2007) A method for the general identification of protein crystals in crystallization experiments using a noncovalent fluorescent dye. Acta Crystallogr D Biol Crystallogr 63:526–535. doi:10.1107/S0907444906056137
Haquin S, Oeuillet E, Pajon A et al (2008) Data management in structural genomics: an overview. Methods Mol Biol 426:49–79. doi:10.1007/978-1-60327-058-8_4
Harp JM, Timm DE, Bunick GJ (1998) Macromolecular crystal annealing: overcoming increased mosaicity associated with cryocrystallography. Acta Crystallogr D Biol Crystallogr 54:622–628
Harp JM, Hanson BL, Timm DE, Bunick GJ (1999) Macromolecular crystal annealing: evaluation of techniques and variables. Acta Crystallogr D Biol Crystallogr 55:1329–1334
Harrill AH, Rusyn I (2008) Systems biology and functional genomics approaches for the identification of cellular responses to drug toxicity. Expert Opin Drug Metab Toxicol 4:1379–1389. doi:10.1517/17425255.4.11.1379
Hartmann H, Parak F, Steigemann W et al (1982) Conformational substates in a protein: structure and dynamics of metmyoglobin at 80 K. Proc Natl Acad Sci USA 79:4967–4971
Heinemann U, Büssow K, Mueller U, Umbach P (2003) Facilities and methods for the high-throughput crystal structural analysis of human proteins. Acc Chem Res 36:157–163
Heras B, Martin JL (2005) Post-crystallization treatments for improving diffraction quality of protein crystals. Acta Crystallogr D Biol Crystallogr 61:1173–1180
Heras B, Edeling MA, Byriel KA et al (2003) Dehydration converts DsbG crystal diffraction from low to high resolution. Structure 11:139–145
Hilgart MC, Sanishvili R, Ogata CM et al (2011) Automated sample-scanning methods for radiation damage mitigation and diffraction-based centering of macromolecular crystals. J Synchrotron Radiat 18:717–722. doi:10.1107/S0909049511029918
Hiraki M, Kato R, Nagai M et al (2006) Development of an automated large-scale protein-crystallization and monitoring system for high-throughput protein-structure analyses. Acta Crystallogr D Biol Crystallogr 62:1058–1065. doi:10.1107/S0907444906023821
Hodis E, Prilusky J, Sussman JL (2010) Proteopedia: A collaborative, virtual 3D web-resource for protein and biomolecule structure and function. Biochem Mol Biol Educ 38:341–342. doi:10.1002/bmb.20431
Holm L, Sander C (1999) Protein folds and families: sequence and structure alignments. Nucleic Acids Res 27:244–247
Holton JM, Frankel KA (2010) The minimum crystal size needed for a complete diffraction data set. Acta Crystallogr D Biol Crystallogr 66:393–408. doi:10.1107/S0907444910007262
Holz C, Prinz B, Bolotina N et al (2003) Establishing the yeast *Saccharomyces cerevisiae* as a system for expression of human proteins on a proteome-scale. J Struct Funct Genomics 4:97–108
Hope H (1988) Cryocrystallography of biological macromolecules: a generally applicable method. Acta Crystallographica Section B Structural Science 44:22–26. doi:10.1107/S0108768187008632
Hope H (1990) Crystallography of biological macromolecules at ultra-low temperature. Annu Rev Biophys Biophys Chem 19:107–126. doi:10.1146/annurev.bb.19.060190.000543
Huang L, Hung L, Odell M et al (2002) Structure-based experimental confirmation of biochemical function to a methyltransferase, MJ0882, from hyperthermophile *Methanococcus jannaschii*. J Struct Funct Genomics 2:121–127
Huang YJ, Hang D, Lu LJ et al (2008) Targeting the human cancer pathway protein interaction network by structural genomics. Mol Cell Proteomics 7:2048–2060. doi:10.1074/mcp.M700550-MCP200
Ishida T, Kinoshita K (2008) Prediction of disordered regions in proteins based on the meta approach. Bioinformatics 24:1344–1348. doi:10.1093/bioinformatics/btn195
Jain D, Lamour V (2010) Computational tools in protein crystallography. Methods Mol Biol 673:129–156. doi:10.1007/978-1-60761-842-3_8
Joachimiak A (2009) High-throughput crystallography for structural genomics. Curr Opin Struct Biol 19:573–584. doi:10.1016/j.sbi.2009.08.002
Jones TA, Zou JY, Cowan SW, Kjeldgaard M (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A 47(Pt 2):110–119
Judge RA, Swift K, Gonzalez C (2005) An ultraviolet fluorescence-based method for identifying and distinguishing protein crystals. Acta Crystallogr D Biol Crystallogr 61:60–66. doi:10.1107/S0907444904026538
Junge F, Haberstock S, Roos C et al (2011) Advances in cell-free protein synthesis for the functional and structural analysis of membrane proteins. New Biotechnology 28:262–271
Kabsch W (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. JApplCryst 26:795–800
Keegan RM, Winn MD (2008) MrBUMP: an automated pipeline for molecular replacement. Acta Crystallogr D Biol Crystallogr 64:119–124. doi:10.1107/S0907444907037195
Kisselman G, Qiu W, Romanov V et al (2011) X-CHIP: an integrated platform for high-throughput protein crystallization and on-the-chip X-ray diffraction data collection. Acta Crystallogr D Biol Crystallogr 67:533–539. doi:10.1107/S0907444911011589
Kissick DJ, Gualtieri EJ, Simpson GJ, Cherezov V (2010) Nonlinear optical imaging of integral membrane protein crystals in lipidic mesophases. Anal Chem 82:491–497. doi:10.1021/ac902139w

Kissick DJ, Wanapun D, Simpson GJ (2011) Second-order nonlinear optical imaging of chiral crystals. Annu Rev Anal Chem (Palo Alto Calif) 4:419–437. doi:10.1146/annurev.anchem.111808.073722

Koide S (2009) Engineering of recombinant crystallization chaperones. Curr Opin Struct Biol 19:449–457. doi:10.1016/j.sbi.2009.04.008

Kriz A, Schmid K, Baumgartner N et al (2010) A plasmid-based multigene expression system for mammalian cells. Nat Commun 1:120. doi:10.1038/ncomms1120

Kummel D, Oeckinghaus A, Wang C et al (2008) Distinct isocomplexes of the TRAPP trafficking factor coexist inside human cells. FEBS Lett 582:3729–3733. doi:10.1016/j.febslet.2008.09.056

Langer G, Cohen SX, Lamzin VS, Perrakis A (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. Nat Protoc 3:1171–1179. doi:10.1038/nprot.2008.91

Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13:323–330, 307–308

Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33:W89–W93

Lavault B, Ravelli RB, Cipriani F (2006) C3D: a program for the automated centring of cryocooled crystals. Acta Crystallogr D Biol Crystallogr 62:1348–1357

Lee JE, Fusco ML, Ollmann Saphire E (2009) An efficient platform for screening expression and crystallization of glycoproteins produced in human cells. Nat Protoc 4:592–604

Lees JA, Yip CK, Walz T, Hughson FM (2010) Molecular organization of the COG vesicle tethering complex. Nat Struct Mol Biol 17:1292–1297. doi:10.1038/nsmb.1917

Leslie AG (2006) The integration of macromolecular diffraction data. Acta Crystallogr D Biol Crystallogr 62:48–57

Listwan P, Terwilliger TC, Waldo GS (2009) Automated, high-throughput platform for protein solubility screening using a split-GFP system. J Struct Funct Genomics 10:47–55. doi:10.1007/s10969-008-9049-4

Liu R, Freund Y, Spraggon G (2008) Image-based crystal detection: a machine-learning approach. Acta Crystallogr D Biol Crystallogr 64:1187–1195. doi:10.1107/S090744490802982X

Long F, Vagin AA, Young P, Murshudov GN (2008) BALBES: a molecular-replacement pipeline. Acta Crystallogr D Biol Crystallogr 64:125–132. doi:10.1107/S0907444907050172

Makino S, Goren MA, Fox BG, Markley JL (2010) Cell-free protein synthesis technology in NMR high-throughput structure determination. Methods Mol Biol 607:127–147. doi:10.1007/978-1-60327-331-2_12

Manjasetty BA, Chance MR (2006) Crystal structure of Escherichia coli L-arabinose isomerase (ECAI), the putative target of biological tagatose production. J Mol Biol 360:297–309. doi:10.1016/j.jmb.2006.04.040

Manjasetty BA, Croteau N, Powlowski J, Vrielink A (2001) Crystallization and preliminary X-ray analysis of dmpFG-encoded 4-hydroxy-2-ketovalerate aldolase-aldehyde dehydrogenase (acylating) from *Pseudomonas* sp. strain CF600. Acta Crystallogr D Biol Crystallogr 57:582–585

Manjasetty BA, Delbruck H, Pham DT et al (2004a) Crystal structure of Homo sapiens protein hp14.5. Proteins 54:797–800

Manjasetty BA, Niesen FH, Delbruck H et al (2004b) X-ray structure of fumarylacetoacetate hydrolase family member Homo sapiens FLJ36880. Biol Chem 385:935–942. doi:10.1515/BC.2004.122

Manjasetty BA, Büssow K, Fieber-Erdmann M et al (2006) Crystal structure of *Homo sapiens* PTD012 reveals a zinc-containing hydrolase fold. Protein Sci 15:914–920. doi:10.1110/ps.052037006

Manjasetty BA, Shi W, Zhan C et al (2007) A high-throughput approach to protein structure analysis. Genet Eng (N Y) 28:105–128

Manjasetty BA, Turnbull AP, Panjikar S et al (2008) Automated technologies and novel techniques to accelerate protein crystallography for structural genomics. Proteomics 8:612–625. doi:10.1002/pmic.200700687

Manjasetty B, Turnbull A, Panjikar S (2010) The impact of structural proteomics on biotechnology. Biotechnol Genet Eng Rev 26:353–370

McCall EJ, Danielsson A, Hardern IM et al (2005) Improvements to the throughput of recombinant protein expression in the baculovirus/insect cell system. Protein Expr Purif 42:29–36

McCoy AJ, Grosse-Kunstleve RW, Adams PD et al (2007) Phaser crystallographic software. J Appl Crystallogr 40:658–674. doi:10.1107/S0021889807021206

McPhillips TM, McPhillips SE, Chiu HJ et al (2002) Blu-Ice and the distributed control system: software for data acquisition and instrument control at macromolecular crystallography beamlines. J Synchrotron Radiat 9:401–406 pii:S0909049502015170

McRee DE, Israel M (2008) XtalView, protein structure solution and protein graphics, a short history. J Struct Biol 163:208–213. doi:10.1016/j.jsb.2008.02.004

Minor W, Cymborowski M, Otwinowski Z, Chruszcz M (2006) HKL-3000: the integration of data reduction and structure solution–from diffraction images to an initial model in minutes. Acta Crystallogr D Biol Crystallogr 62:859–866. doi:10.1107/S0907444906019949

Moon AF, Mueller GA, Zhong X, Pedersen LC (2010) A synergistic approach to protein crystallization: combination of a fixed-arm carrier with surface entropy reduction. Protein Sci 19:901–913. doi:10.1002/pro.368

Morris RJ, Perrakis A, Lamzin VS (2003) ARP/wARP and automatic interpretation of protein electron density maps. Methods Enzymol 374:229–244

Morris C, Pajon A, Griffiths SL et al (2011) The protein information management system (PiMS): a generic tool for any structural biology research laboratory. Acta Crystallogr D Biol Crystallogr 67:249–260. doi:10.1107/S0907444911007943

Mueller-Dieckmann C, Kauffmann B, Weiss MS (2011) Trimethylamine N-oxide as a versatile cryoprotective agent in macromolecular crystallography. J Appl Crystallogr 44:433–436. doi:10.1107/S0021889811000045

Munoz-Elias EJ, McKinney JD (2005) Mycobacterium tuberculosis isocitrate lyases 1 and 2 are jointly required for in vivo growth and virulence. Nat Med 11:638–644. doi:10.1038/nm1252

Musa TL, Ioerger TR, Sacchettini JC (2009) The tuberculosis structural genomics consortium: a structural genomics approach to drug discovery. Adv Protein Chem Struct Biol 77:41–76. doi:10.1016/S1876-1623(09)77003-8

Naistat DM, Leblanc R (2004) Proteomics. J Environ Pathol Toxicol Oncol 23:161–178

Nanao MH, Ravelli RB (2006) Phasing macromolecular structures with UV-induced structural changes. Structure 14:791–800. doi:10.1016/j.str.2006.02.007

Navaza J (2001) Implementation of molecular replacement in AMoRe. Acta Crystallogr D Biol Crystallogr 57:1367–1372

Ness SR, de Graaff RA, Abrahams JP, Pannu NS (2004) CRANK: new methods for automated macromolecular crystal structure solution. Structure (Camb) 12:1753–1761

Nie Y, Viola C, Bieniossek C et al (2009) Getting a grip on complexes. Curr Genomics 10:558–572. doi:10.2174/138920209789503923

Ostheimer GJ, Barkan A, Matthews BW (2002) Crystal structure of *E. coli* YhbY: a representative of a novel class of RNA binding proteins. Structure 10:1593–1601 pii:S0969212602008869

Otwinowski Z, Minor W (1997) Processing of x-ray diffraction data collected in oscillation mode. In: Carter CW Jr, Sweet RM (eds) Methods in Enzymology. Academic Press, New York, pp 307–326

Paithankar KS, Garman EF (2010) Know your dose: RADDOSE. Acta Crystallogr D Biol Crystallogr 66:381–388. doi:10.1107/S0907444910006724

Pan X, Wesenberg GE, Markley JL et al (2007) A graphical approach to tracking and reporting target status in structural genomics. J Struct Funct Genomics 8:209–216. doi:10.1007/s10969-007-9037-0

Panjikar S, Parthasarathy V, Lamzin VS et al (2005) Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. Acta Crystallogr D Biol Crystallogr 61:449–457

Panjikar S, Parthasarathy V, Lamzin VS et al (2009) On the combination of molecular replacement and single-wavelength anomalous diffraction phasing for automated structure determination. Acta Crystallogr D Biol Crystallogr 65:1089–1097. doi:10.1107/S0907444909029643

Panjikar S, Mayerhofer H, Tucker PA et al (2011) Single isomorphous replacement phasing of selenomethionine-containing proteins using UV-induced radiation damage. Acta Crystallogr D Biol Crystallogr 67:32–44. doi:10.1107/S090744491004299X

Pannu NS, Read RJ (2004) The application of multivariate statistical techniques improves single-wavelength anomalous diffraction phasing. Acta Crystallogr D Biol Crystallogr 60:22–27 pii:S0907444903020808

Pannu NS, Waterreus WJ, Skubak P et al (2011) Recent advances in the CRANK software suite for experimental phasing. Acta Crystallogr D Biol Crystallogr 67:331–337. doi:10.1107/S0907444910052224

Pape T, Schneider TR (2004) HKL2MAP: a graphical user interface for phasing with SHELX programs. J Appl Cryst 37:843–844

Pedelacq JD, Nguyen HB, Cabantous S et al (2011) Experimental mapping of soluble protein domains using a hierarchical approach. Nucleic Acids Res. doi:10.1093/nar/gkr548

Perrakis A, Musacchio A, Cusack S, Petosa C (2011) Investigating a macromolecular complex: the toolkit of methods. J Struct Biol 175:106–112. doi:10.1016/j.jsb.2011.05.014

Petrek M, Otyepka M, Banas P et al (2006) CAVER: a new tool to explore routes from protein clefts, pockets and cavities. BMC Bioinformatics 7:316. doi:10.1186/1471-2105-7-316

Petschnigg J, Snider J, Stagljar I (2011) Interactive proteomics research technologies: recent applications and advances. Curr Opin Biotechnol 22:50–58. doi:10.1016/j.copbio.2010.09.001

Pike AC, Rellos P, Niesen FH, Turnbull A, Oliver AW, Parker SA, Turk BE, Pearl LH, Knapp S (2008) Activation segment dimerization: a mechanism for kinase autophosphorylation of non-consensus sites. EMBO J 27(4):704–714. doi:10.1038/emboj.2008.8

Pothineni SB, Strutz T, Lamzin VS (2006) Automated detection and centring of cryocooled protein crystals. Acta Crystallogr D Biol Crystallogr 62:1358–1368. doi:10.1107/S0907444906031672

Prilusky J, Oueillet E, Ulryck N et al (2005) HalX: an open-source LIMS (laboratory information management system) for small- to large-scale laboratories. Acta Crystallogr D Biol Crystallogr 61:671–678. doi:10.1107/S0907444905001290

Prilusky J, Hodis E, Canner D et al (2011) Proteopedia: a status report on the collaborative, 3D web-encyclopedia of proteins and other biomolecules. J Struct Biol 175:244–252. doi:10.1016/j.jsb.2011.04.011

Raimondo D, Giorgetti A, Bosi S, Tramontano A (2007) Automatic procedure for using models of proteins in molecular replacement. Proteins 66:689–696. doi:10.1002/prot.21225

Rasmussen SG, Choi HJ, Rosenbaum DM et al (2007) Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. Nature 450:383–387. doi:10.1038/nature06325

Rasmussen SG, Choi HJ, Fung JJ et al (2011) Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor. Nature 469:175–180. doi:10.1038/nature09648

Raunest M, Kandt C (2011) dxTuber: detecting protein cavities, tunnels and clefts based on protein and solvent dynamics. J Mol Graph Model 29:895–905. doi:10.1016/j.jmgm.2011.02.003

Ravelli RB, Leiros HK, Pan B, Caffrey M, McSweeney S (2003) Shedding UV light on phase problem. Structure 11:217–224. doi:10.1016/j.str.2006.03.004

Raymond S, O'Toole N, Cygler M (2004) A data management system for structural genomics. Proteome Sci 2:4

Reckel S, Sobhanifar S, Durst F et al (2010) Strategies for the cell-free expression of membrane proteins. Methods Mol Biol 607:187–212. doi:10.1007/978-1-60327-331-2_16

Reddy V, Swanson SM, Segelke B et al (2003) Effective electron-density map improvement and structure validation on a Linux multi-CPU web cluster: the TB structural genomics consortium bias removal web service. Acta Crystallogr D Biol Crystallogr 59:2200–2210 pii:S0907444903020316

Reeves PJ, Callewaert N, Contreras R, Khorana HG (2002) Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. Proc Natl Acad Sci USA 99:13419–13424. doi:10.1073/pnas.212519299.212519299

Rellos P, Pike AC, Niesen FH et al (2010) Structure of the CaMKIIdelta/calmodulin complex reveals the molecular mechanism of CaMKII kinase activation. PLoS Biol 8:e1000426. doi:10.1371/journal.pbio.1000426

Ren Y, Yip CK, Tripathi A et al (2009) A structure-based mechanism for vesicle capture by the multisubunit tethering complex Dsl1. Cell 139:1119–1129. doi:10.1016/j.cell.2009.11.002

Riboldi-Tunnicliffe A, Hilgenfeld R (1999) Cryocrystallography with oil–an old idea revived. J Appl Crystallogr 32:1003–1005. doi:10.1107/S0021889899008584

Rodriguez DD, Grosse C, Himmel S et al (2009) Crystallographic ab initio protein structure solution below atomic resolution. Nat Methods 6:651–653. doi:10.1038/nmeth.1365

Russi S, Juers DH, Sanchez-Weatherby J et al (2011) Inducing phase changes in crystals of macromolecules: Status and perspectives for controlled crystal dehydration. J Struct Biol 175:236–243. doi:10.1016/j.jsb.2011.03.002

Samygina VR, Antonyuk SV, Lamzin VS, Popov AN (2000) Improving the X-ray resolution by reversible flash-cooling combined with concentration screening, as exemplified with PPase. Acta Crystallogr D Biol Crystallogr 56:595–603 (HE0242)

Sanchez-Weatherby J, Bowler MW, Huet J et al (2009) Improving diffraction by humidity control: a novel device compatible with X-ray beamlines. Acta Crystallogr D Biol Crystallogr 65:1237–1246. doi:10.1107/S0907444909037822

Sasson O, Linial M (2005) ProTarget: automatic prediction of protein structure novelty. Nucleic Acids Res 33:W81–W84

Scheich C, Leitner D, Sievert V et al (2004) Fast identification of folded human protein domains expressed in *E. coli* suitable for structural analysis. BMC Struct Biol 4:4

Scheich C, Kümmel D, Soumailakakis D et al (2007) Vectors for co-expression of an unrestricted number of proteins. Nucleic Acids Res 35:e43. doi:10.1093/nar/gkm067

Schneider TR, Sheldrick GM (2002) Substructure solution with SHELXD. Acta Crystallogr D Biol Crystallogr 58:1772–1779

Schuermann JP, Tanner JJ (2003) MRSAD: using anomalous dispersion from S atoms collected at Cu Kalpha wavelength in

molecular-replacement structure determination. Acta Crystallogr D Biol Crystallogr 59:1731–1736 pii:S0907444903015725
Schulze-Gahmen U, Pelaschier J, Yokota H et al (2003) Crystal structure of a hypothetical protein, TM841 of *Thermotoga maritima*, reveals its function as a fatty acid-binding protein. Proteins 50:526–530
Schwarzenbacher R, Godzik A, Jaroszewski L (2008) The JCSG MR pipeline: optimized alignments, multiple models and parallel searches. Acta Crystallogr D Biol Crystallogr 64:133–140. doi:10.1107/S0907444907050111
Segelke B (2005) Macromolecular crystallization with microfluidic free-interface diffusion. Expert Rev Proteomics 2:165–172
Sharma V, Sharma S, Hoener Zu, Bentrup K, McKinney JD, Russell DG, Jacobs WR Jr, Sacchettini JC (2000) Structure of isocitrate lyase, a persistence factor of Mycobacterium tuberculosis. Nat Struct Biol 7:663–668. doi:10.1038/77964
Shi W, Zhan C, Ignatov A et al (2005) Metalloproteomics: high-throughput structural and functional annotation of proteins in structural genomics. Structure (Camb) 13:1473–1486
Shi W, Punta M, Bohon J et al (2011) Characterization of metalloproteins by high-throughput X-ray absorption spectroscopy. Genome Res 21:898–907. doi:10.1101/gr.115097.110
Shimamura T, Shiroishi M, Weyand S et al (2011) Structure of the human histamine H1 receptor complex with doxepin. Nature 475:65–70. doi:10.1038/nature10236
Shin DH, Hou J, Chandonia JM et al (2007) Structure-based inference of molecular functions of proteins of unknown function from Berkeley Structural Genomics Center. J Struct Funct Genomics 8:99–105. doi:10.1007/s10969-007-9025-4
Skinner JM, Cowan M, Buono R et al (2006) Integrated software for macromolecular crystallography synchrotron beamlines II: revision, robots and a database. Acta Crystallogr D Biol Crystallogr 62:1340–1347. doi:10.1107/S0907444906030162
Sledz P, Zheng H, Murzyn K et al (2010) New surface contacts formed upon reductive lysine methylation: improving the probability of protein crystallization. Protein Sci 19:1395–1404. doi:10.1002/pro.420
Snell G, Cork C, Nordmeyer R et al (2004) Automated sample mounting and alignment system for biological crystallography at a synchrotron source. Structure 12:537–545
Soares AS, Engel MA, Stearns R et al (2011) Acoustically mounted microcrystals yield high-resolution X-ray structures. Biochemistry 50:4399–4401. doi:10.1021/bi200549x
Song J, Mathew D, Jacob SA et al (2007) Diffraction-based automated crystal centering. J Synchrotron Radiat 14:191–195. doi:10.1107/S0909049507004803
Standfuss J, Xie G, Edwards PC et al (2007) Crystal structure of a thermally stable rhodopsin mutant. J Mol Biol 372:1179–1188. doi:10.1016/j.jmb.2007.03.007
Standfuss J, Edwards PC, D'Antona A et al (2011) The structural basis of agonist-induced activation in constitutively active rhodopsin. Nature 471:656–660. doi:10.1038/nature09795
Stanley P (1989) Chinese hamster ovary cell mutants with multiple glycosylation defects for production of glycoproteins with minimal carbohydrate heterogeneity. Mol Cell Biol 9:377–383
Stark A, Russell RB (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. Nucleic Acids Res 31:3341–3344
Stepanov S, Makarov O, Hilgart M et al (2011) JBluIce-EPICS control system for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 67:176–188. doi:10.1107/S0907444910053916
Stojanoff V, Jakoncic J, Oren DA et al (2011) From screen to structure with a harvestable microfluidic device. Acta Crystallogr Sect F Struct Biol Cryst Commun 67:971–975. doi:10.1107/S1744309111024456
Strokopytov BV, Fedorov A, Mahoney NM et al (2005) Phased translation function revisited: structure solution of the cofilin-homology domain from yeast actin-binding protein 1 using six-dimensional searches. Acta Crystallogr D Biol Crystallogr 61:285–293
Stults JT, Arnott D (2005) Proteomics. Methods Enzymol 402:245–289
Teplyakov A, Obmolova G, Khil PP et al (2003) Crystal structure of the *Escherichia coli* YcdX protein reveals a trinuclear zinc active site. Proteins 51:315–318. doi:10.1002/prot.10352
Terwilliger TC (2000) Maximum-likelihood density modification. Acta Crystallogr D Biol Crystallogr 56:965–972 pii:S0907444900005072
Terwilliger TC (2011) The success of structural genomics. J Struct Funct Genomics 12:43–44. doi:10.1007/s10969-011-9114-2
Thomanek UF, Parak F, Mosbauer RL et al (1973) Freezing of myoglobin crystals at high pressure. Acta Crystallogr A 29:263–265
Trowitzsch S, Bieniossek C, Nie Y et al (2010) New baculovirus expression tools for recombinant protein complex production. J Struct Biol 172:45–54. doi:10.1016/j.jsb.2010.02.010
Uysal S, Vasquez V, Tereshko V et al (2009) Crystal structure of full-length KcsA in its closed conformation. Proc Natl Acad Sci USA 106:6644–6649. doi:10.1073/pnas.0810663106
Vagin A, Teplyakov A (1997) MOLREP: an automated program for molecular replacement. JApplCryst 30:1022–1025
Vahedi-Faridi A, Stojanoff V, Yeh JI (2005) The effects of flash-annealing on glycerol kinase crystals. Acta Crystallogr D Biol Crystallogr 61:982–989. doi:10.1107/S0907444905012746
Vallejo LF, Rinas U (2004) Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. Microb Cell Fact 3:11. doi:10.1186/1475-2859-3-11
Vijayachandran LS, Viola C, Garzoni F et al (2011) Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. J Struct Biol 175:198–208. doi:10.1016/j.jsb.2011.03.007
Vincentelli R, Canaan S, Campanacci V et al (2004) High-throughput automated refolding screening of inclusion bodies. Protein Sci 13:2782–2792. doi:10.1110/ps.04806004
von Grotthuss M, Plewczynski D, Vriend G, Rychlewski L (2008) 3D-Fun: predicting enzyme function from structure. Nucleic Acids Res 36:W303–W307. doi:10.1093/nar/gkn308
Vonrhein C, Blanc E, Roversi P, Bricogne G (2007) Automated structure solution with autoSHARP. Methods Mol Biol 364:215–230. doi:10.1385/1-59745-266-1:215
Walter TS, Diprose JM, Mayo CJ et al (2005) A procedure for setting up high-throughput nanolitre crystallization experiments Crystallization workflow for initial screening, automated storage, imaging and optimization. Acta Crystallogr D Biol Crystallogr 61:651–657. doi:10.1107/S0907444905007808
Walter TS, Meier C, Assenberg R et al (2006) Lysine methylation as a routine rescue strategy for protein crystallization. Structure 14:1617–1622. doi:10.1016/j.str.2006.09.005
Walter TS, Mancini EJ, Kadlec J et al (2008) Semi-automated microseeding of nanolitre crystallization experiments. Acta Crystallogr Sect F Struct Biol Cryst Commun 64:14–18. doi:10.1107/S1744309107057260
Ward JJ, Sodhi JS, McGuffin LJ et al (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337:635–645. doi:10.1016/j.jmb.2004.02.002
Watanabe M, Miyazono K, Tanokura M et al (2010) Cell-free protein synthesis for structure determination by X-ray crystallography. Methods Mol Biol 607:149–160. doi:10.1007/978-1-60327-331-2_13
Weekes D, Krishna SS, Bakolitsa C et al (2010) TOPSAN: a collaborative annotation environment for structural genomics. BMC Bioinformatics 11:426. doi:10.1186/1471-2105-11-426

Weeks CM, Blessing RH, Miller R et al (2002) Towards automated protein structure determination: BnP, the SnB-PHASES interface. Z Kristallogr 217:686–693

Wilke S, Krausze J, Gossen M et al (2010) Glycoprotein production for structure analysis with stable, glycosylation mutant CHO cell lines established by fluorescence-activated cell sorting. Protein Sci 19:1264–1271

Wilke S, Groebe L, Maffenbeier V, et al (2011) Streamlining Homogeneous Glycoprotein Production for Biophysical and Structural Applications by Targeted Cell Line Development. PLoS ONE. (in press)

Winn MD, Ballard CC, Cowtan KD et al (2011) Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr 67:235–242. doi:10.1107/S0907444910045749

Xu H, Weeks CM (2008) Rapid and automated substructure solution by Shake-and-Bake. Acta Crystallogr D Biol Crystallogr 64:172–177. doi:10.1107/S0907444907058659

Yakunin AF, Yee AA, Savchenko A et al (2004) Structural proteomics: a tool for genome annotation. Curr Opin Chem Biol 8:42–48

Yang ZR, Thomson R, McNeil P, Esnouf RM (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21:3369–3376. doi:10.1093/bioinformatics/bti534

Yeh JI, Hol WG (1998) A flash-annealing technique to improve diffraction limits and lower mosaicity in crystals of glycerol kinase. Acta Crystallogr D Biol Crystallogr 54:479–480

Yumerefendi H, Desravines DC, Hart DJ (2011) Library-based methods for identification of soluble expression constructs. Methods. doi:10.1016/j.ymeth.2011.06.007

Zhan C, Fedorov EV, Shi W et al (2005) The ybeY protein from *Escherichia coli* is a metalloprotein. Acta Crystallogr Sect F Struct Biol Cryst Commun 61:959–963. doi:10.1107/S1744309105031131

Zolnai Z, Lee PT, Li J et al (2003) Project management system for structural and functional proteomics: Sesame. J Struct Funct Genomics 4:11–23