1
Advances in High-throughput Methodologies for Crystallizing Proteins

P.P. CHAMBERLAIN[1], H.E. KLOCK[1,2], D. McMULLAN[1,2], M. DiDONATO[1,2], A. KREUSCH[1,2], S.A. LESLEY[1,2] AND G. SPRAGGON[1,2]*

[1]*Genomics Institute of the Novartis Research Foundation and* [2]*Joint Center for Structural Genomics, Crystallomics Core, 10675 John Jay Hopkins Drive, San Diego, CA 92121, USA*

Introduction

Following the success of the human genome project, the Protein Structure Initiative (PSI) was started by the National Institutes of Health (NIH) to provide high-throughput determination of protein structures by either X-ray crystallography or nuclear magnetic resonance (NMR). The goal of the project is to provide the means to map out protein fold space, thus making all translated genes in the various sequenced genomes accessible to molecular modelling (Burley, 2000). The three-dimensional information derived can be used for applications ranging from studying the evolution and biochemistry of proteins to the design of small molecule inhibitors for the treatment of disease. Compared with genomic efforts, structural genomics faces significant challenges since the physicochemical parameters and behaviour of proteins vary widely. Traditionally, this has led to highly customized processes applicable only to well-characterized, high-value targets. The development of generic approaches for the expression, purification, and crystallization of poorly characterized proteins has been vital therefore for the successful parallelization required for the structural genomic approach.

*To whom correspondence may be addressed (gspraggo@gnf.org)

Abbreviations: ASU, asymmetric unit; BMCD, Biological Macromolecular Crystallization Database; BU, biological unit; CEEP, Crystal Experiment Evaluation Program; DLS, dynamic light scattering; DXMS, deuterium exchange mass spectroscopy; GNF, Genomics Institute of the Novartis Research Foundation; GUI, graphical user interface; HPLC, high pressure liquid chromatography; IPTG, isopropyl-β-D-thiogalactopyranoside; JCSG, Joint Center for Structural Genomics; MAD, multiwavelength anomalous dispersion; MIR, multiple isomorphous replacement; MW, molecular weight; NIH, National Institutes of Health; NMR, nuclear magnetic resonance; ORF, open reading frame; PCR, polymerase chain reaction; PDB, Protein Data Bank; PEG, polyethylene glycol; PIMS, Protein Information Management System; PQS, protein quaternary structure server; PSI, Protein Structure Initiative; SEC, size exclusion chromatography; SLS, static light scattering; TBS, tris-buffered saline.

Biotechnology and Genetic Engineering Reviews – Vol. 23, December 2006
0264-8725/07/23/1-19 $20.00 + $0.00 © Lavoisier/Intercept, 14 rue de Provigny, F-94236 Cachan cedex, France

Initially, nine structural genomics centres were funded for a period of five years. This, coupled with the establishment of global sister initiatives and a number of structural genomics companies distributed around the world (Stevens and Wilson, 2001), has provided a great impetus for the development of high-throughput and automated systems. Recent years have witnessed the appearance of automated systems for parallel expression and purification, robotic crystallization and imaging, and remote collection of X-ray data (Abola *et al.*, 2000; DiDonato *et al.*, 2004). Since the beginning of the structural genomics era, around a quarter of *de novo* structures deposited in the Protein Data Bank (www.rcbs.org/pdb/) were produced by structural genomics initiatives (Terwilliger, 2004). Phase Two of the PSI is now under way, comprising four large structural genomics centres and six additional specialized centres.

In the structure solution process, protein crystallization is often described as the principal bottleneck, although this point is increasingly debated. Protein crystallization is a process whereby a protein solution is brought into an ordered solid phase by inducing a state of protein supersaturation (Haas and Drenth, 1999). Various methods exist for achieving a supersaturated state but, in general, either vapour diffusion or liquid diffusion techniques are used, which involve equilibration of the protein droplet against a solution of different solute potential.

Most crystallization screening methods depend on exposing a high concentration of protein (typically >10 mg/ml) to a variety of salts, organic molecules, and polymers to identify the specific conditions required to induce nucleation of crystals. Crystallization of proteins can be extremely difficult and laborious, largely because there are no methods available for predicting conditions under which good protein crystals will form.

Due to the number of variables and the wide range of conditions under which crystals may occur, a full factorial screen would be impractical. An influential innovation was the introduction to crystallography of sparse matrix methods (Jancarik and Kim, 1991). Initially, a random screen of fifty conditions, dubbed the 'Magic 50' and shown to have produced protein crystals in the past, was introduced as a screening method for crystal growth. The range and variety of crystal screens have continued to develop to such an extent that they now typically consist of hundreds of conditions, and the accumulation of experimental crystallization data should mean that screens will continue to evolve.

The sparse matrix screening strategy is costly in terms of material, the most precious of which is likely to be the highly purified and concentrated protein. The high attrition rates during screening place an onus on the development of methods capable of reducing the demands on material whilst retaining coverage. An important advance in sparse matrix screening was the miniaturization and automation of the crystallization experiments (Santarsiero *et al.*, 2002). Based on a robot, 'the T2K', designed at the Lawrence Berkeley lab (Stevens, 2000), the original nanovolume high-throughput system, named Agincourt, was developed (Hosfield *et al.*, 2003). Droplet volumes were typically 100 nl, thus reducing the demand for protein by around a factor of ten. Not only does miniaturization reduce the need for materials, it also speeds up the equilibration process, and thus the production of crystals (Santarsiero *et al.*, 2002).

The submicrolitre vapour diffusion format is fast becoming a standard within the

structural biology community, and is well established in a number of laboratories (Brown *et al.*, 2003), as well as industrial concerns (Hosfield *et al.*, 2003), although alternatives such as microbatch are also proving successful (Luft *et al.*, 1999, 2003). It remains to be seen how much further experimental volumes can be reduced while still yielding usable crystals. The increasing drive to miniaturize experiments has led to experimental apparatus incorporating microfluidic 'lab on a chip' technology (Hansen *et al.*, 2002; Hansen and Quake, 2003). Such systems use interface–interface diffusion and hydraulics for crystallization and can perform hundreds of experiments with 10 µl of protein. Very small crystals present challenges in handling (crystals remain to be transferred to loops by hand, prior to cooling) and also scatter X-rays more weakly. It is possible that further advances in automated or robot-assisted crystal manipulation or *in situ* diffraction methods may help push back the boundary of what constitutes a usable crystal (Abola *et al.*, 2000).

Diffraction data must be collected from the crystal via exposure to hard X-rays (typically on the order of 1.0 Å wavelength). Modern third-generation synchrotron X-ray sources (Helliwell, 1998) are now thousands of times more intense and less divergent than the original apparatus used in X-ray studies. This performance increase provides significant improvement in terms of both the quality of X-ray data and the time needed to collect it, as well as permitting the use of increasingly smaller, poorer crystals (Morris *et al.*, 1999).

In the course of a protein structure determination experiment, diffracted X-ray intensities are measured, but phase information from the diffracted waves is lost. This information must be reconstituted if the structure is to be solved. Where a homologous structure exists, it is possible to solve a structure by calculating the phases from the homologous model computationally placed in the new crystal, a process known as molecular replacement (Hoppe, 1957; Rossmann and Blow, 1962).

However, where no homologous model exists or when molecular replacement is unsuccessful, experimental phasing is required. Common methods for experimentally recovering phase information involve placing heavy atoms within the crystal to provide a reference 'substructure' to determine the phases, a process known as multiple isomorphous replacement (MIR) (Green *et al.*, 1954). MIR was followed some years later by multiwavelength anomalous dispersion (MAD) (Hendrickson, 1991), a method that can efficiently use high-powered tunable X-ray beams from synchrotrons to determine the phases based on the anomalous scattering behaviour of some atoms. The anomalous dispersion method was greatly enhanced by the development of selenomethionine (SeMet) derivation of proteins (Hendrickson *et al.*, 1990), which has been shown to be a reliable and generic way of obtaining phase information (Hendrickson, 1999).

Establishing a high-throughput system for structural biology is a complex multivariate process, which not only has to utilize robust protocols that work for the majority of cases, but also has to be flexible enough to incorporate different methodologies where these procedures do not work or when new technologies become available. Because of the almost infinite possibilities in the process, it is unlikely that all projects will ever be fully automated. In recent years, numerous advances in technology have altered the esoteric, labour-intensive process of crystallization into a more streamlined and rational system.

The shotgun approach to structural genomics often has been described as

'harvesting the low-hanging fruit' of easily crystallizable targets. This assessment may have had some validity early in the PSI; however, current efforts apply these high-throughput, parallelized systems to the development of additional techniques for the solution of recalcitrant proteins. New biophysical methods for enhancing the probability of crystallization are required, and these must be compatible with the existing high-throughput frameworks. Here, we describe some of the progress in such techniques that are being developed and incorporated into the high-throughput pipeline in the crystallomics core of the Joint Center for Structural Genomics (JCSG). The review is roughly split into three sections: a main pipeline through which the targets are submitted to assess their crystallizability, which has been described in previous papers (Lesley *et al.*, 2002; DiDonato *et al.*, 2004); the laboratory information management and imaging systems that are vital to monitoring individual targets and evaluating the process; and finally, the recently developed 'salvage' pathways that are coming on-line for proteins that fail during the standard crystallization or phasing procedures.

GNF/JCSG gene to crystal pipeline

The Genomics Institute of the Novartis Research Foundation (GNF/JCSG) pipeline has been discussed extensively elsewhere (Lesley, 2001; Lesley *et al.*, 2002; DiDonato *et al.*, 2004) but will be described briefly here to provide some context for subsequent discussion.

CLONING

Although options exist for modular, recombination-driven cloning, inclusion of terminal translated sequences by this method in part led to the development of high-throughput conventional cloning methods by the JCSG. The pMH4 expression vector is used for direct ligation of PCR products, with no intermediate subcloning step. The primary vector utilized in expression studies was modified from the pBad/Thio vector (Invitrogen, Inc., USA). The pMH4 vector contains both arabinose-inducible and T7 promoters to allow induction with either IPTG or arabinose, providing tight transcriptional repression and greater flexibility in the selection of cell lines. Translational initiation occurs with the expression/purification tag MGSDKIHHHHHH, which is incorporated at the N-terminus of the translated region. The importance of the first residues following the initiating methionine for final yield have been well determined; in pMH4, the first six residues are derived from thioredoxin and are included for the purpose of increasing and homogenizing expression levels (DiDonato *et al.*, 2004).

EXPRESSION

Crystal screening of several hundred conditions requires a quantity of protein typically in the range of 10 mg, placing a heavy demand on the initial expression. The need for large protein quantities is exacerbated by the requirement for rigorous purification of the target proteins, resulting in inevitable losses during processing. Customized fermentation equipment was developed at GNF capable of running 96

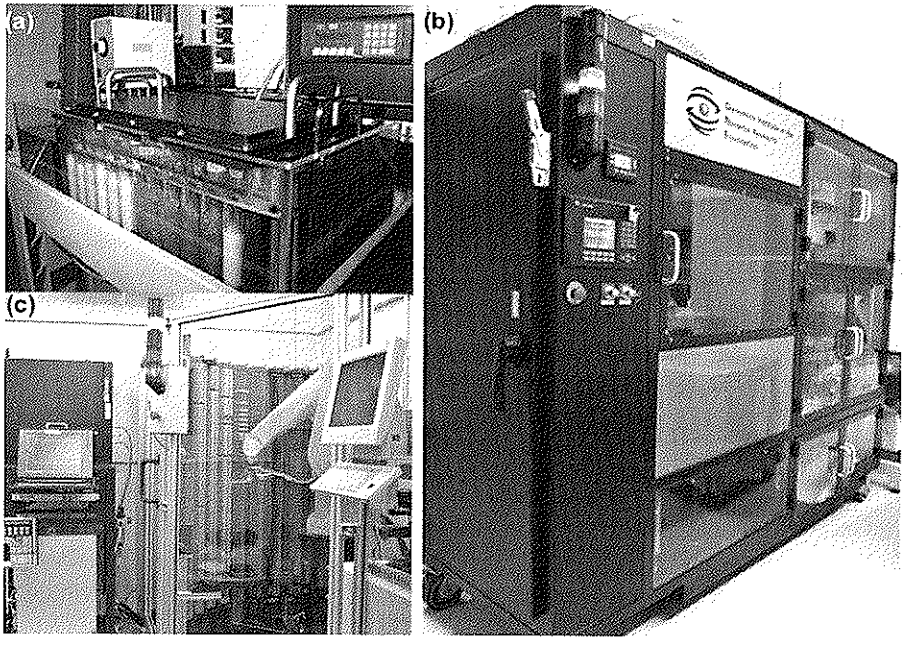

Figure 1.1. The GNF crystallomics technology platform. (a) The fermenter, capable of culturing 96 parallel 65 ml *Escherichia coli* cultures to cell densities of 20–40 OD_{600}. (b) The GNFuge3 possesses liquid handling, sonication, and centrifugation capabilities, used for harvesting and initial purification of recombinant proteins. (c) Crystal imager and plate storage vault. Plates are moved between the racked vault and the imager by the Stäubli arm.

individual 65 ml cultures (*Figure 1.1a*). The reduced volume of the fermentation tubes, when compared to typical shaker flask cultures, is compensated for by the high cell densities achievable in this system (20–40 OD_{600nm}). Air and oxygen sparging makes these cell densities possible, with each tube yielding 2–4 g of bacterial cells, comparable to a 1 litre shaker flask culture. Using the fermentor, 96 cultures can be processed every six hours. Of the *Thermotoga maritima* targets that produce soluble protein, an average of >6.5 mg of protein was isolated per tube after the initial affinity purification step (Lesley *et al.*, 2002; Hosfield *et al.*, 2003).

PURIFICATION

Generalized protocols are a requirement for rapid and parallel protein purification. Since the biophysical properties of proteins vary so widely, this is most readily achieved through the use of affinity purification tags fused to the translated protein. The broad array of affinity tags available has been reviewed elsewhere (Lesley, 2001; Lichty *et al.*, 2005), but the 6-His nickel affinity tag has been adopted by the JCSG and most other structural genomics centres. This tag binds tightly to numerous and inexpensive immobilized metal resins, is highly selective, and is small enough to not interfere significantly with subsequent crystallization. The extraction and harvesting of recombinant proteins is performed using a purpose-developed robot (Lesley, 2001; DiDonato *et al.*, 2004), the GNFuge 3 (*Figure 1.1b*). This refrigerated,

automated extraction and purification system incorporates centrifugal, dispensing, aspirating, sonicating, and affinity purification capabilities. Up to 96 samples can be accommodated per run, with up to five unique buffers applicable while in extraction mode. The system is also compatible with gravity-fed affinity columns in a 96-well block format, with capacity for up to ten unique solutions available in purification format. Proteins tend to require more customized treatment during secondary purification due to the variation in properties between targets. An array of liquid chromatography techniques is available for subsequent purification. Typically, the purification applies anion exchange chromatography, followed by size-exclusion chromatography (SEC) (McMullan *et al.*, 2005).

CRYSTALLIZATION

All crystallization experiments are performed with the Hydra Plus One (Krupka *et al.*, 2002), which was developed itself for structural genomics applications. The Hydra Plus One is capable of performing 96-well format, sitting drop, vapour diffusion experiments in a matter of minutes. Although not as flexible as machines such as Agincourt (Hosfield *et al.*, 2003) and requiring some manual operation, the Hydra system provides a laboratory-based approach to performing experiments with similar specifications. Typical experiments are performed using volumes of protein and crystallant of 250 nl with 50 µl of crystallization solution in the reservoir. Screening is performed with 480 conditions based on commercial sparse matrix designs and grid screens at both 4 and 20°C crystallization temperatures.

Protein information management system

The large number of experiments carried out in high-throughput systems necessitates the use of robust systems to manage and track the complex array of processes and projects, and a number of computational tools have been developed at GNF to meet these demands. The collection of large bodies of consistent experimental data has also presented an opportunity for statistical analysis of crystallographic protocols that previously have been treated as purely trial and error. Although public databases such as the PDB (Berman *et al.*, 2002) and the Biological Macromolecular Crystallization Database (BMCD) (Gilliland *et al.*, 1996) do collect data on crystallization conditions, this information suffers due to the omission of negative results and a number of different recording formats. Central to the collection of data is the JCSG/GNF Protein Information Management System (PIMS). The PIMS is managed via a MySQL database connected to a custom-built Microsoft ACCESS front end (*Figure 1.2*). The ACCESS front end provides a simple interface for querying and entering data into all tables within the database, as well as the ability to display data such as gels, chromatograms, and crystallization images. Of particular use is the ability to recall and display images from all crystallization experiments (*Figure 1.2a*) while displaying the results of annotation. A further view can also assemble results from 'like' conditions in crystallization space (*Figure 1.2b*). This computational re-racking provides a filter that can distinguish optimal pH, additives, and precipitants, thus allowing correlations within crystallization space to be observed. A number of custom applications also connect to the database and communicate data to

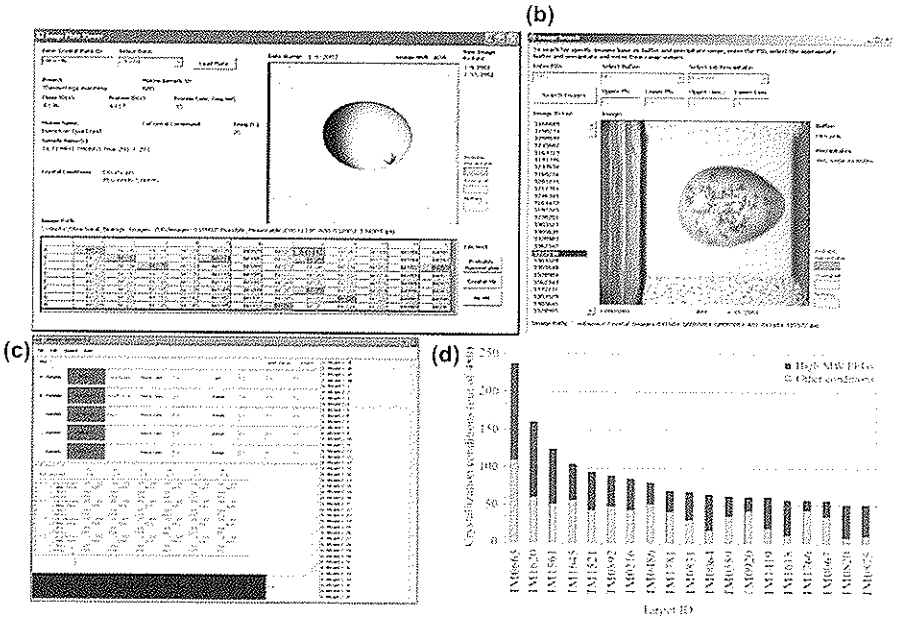

Figure 1.2. The Protein Information Management System (PIMS), a central MySQL database that enables connectivity to a range of applications. (a) The ACCESS front end to the database showing the crystallization plate viewer. (b) The ACCESS front end showing a generalized search of crystallization conditions for comparison of similar conditions. Required buffer formulations are calculated and displayed. (c) NUDGE, a platform independent fine screen generation system with GUI. After a successful crystal hit is identified from screening, NUDGE can be applied to calculate a grid screen focused on this condition. (d) Data mining from the database: a diagram of the number of crystal hits per protein.

collaborators, image viewers, and crystallization fine screen generators (*Figure 1.2c*). Of course, not only is the capturing of data important as a tracking system, it is also useful for data mining. By targeting every open reading frame (ORF) in *T. maritima*, unbiased data for the analysis of patterns in protein behaviour are produced. For example, the resultant data generated from the crystallization of the *T. maritima* proteome has been used subsequently to improve future crystallization experiments (Page *et al.*, 2003) and determine trends within crystallization parameter space (*Figure 1.2d*) (Lesley *et al.*, 2002; Page and Stevens, 2004). Interestingly, these showed that there is no correlation between protein predicted pI and crystallization pH, although differences were identified in crystallization propensity between targets clustered in two dimensions (GRAVY hydropathy index and pI) (Canaves *et al.*, 2004). Successful crystallization results also tended to occur more frequently for proteins with lower molecular weight values and slightly more acidic pHs than the respective *T. maritima* genome-wide averages (Lesley *et al.*, 2002).

IMAGING TECHNOLOGIES

As in all high-throughput systems, the creation of a large number of experiments leads to a high volume of results to inspect, a problem exacerbated by the fact that

good crystallization results occur rarely for most proteins. Due to the large number of experiments performed in typical sparse matrix screens, the automated collection of images and subsequent storage is critical to the success of high-throughput methodologies in structural biology. For example, the first five years of the JCSG resulted in over six million crystallization images.

PLATE STORAGE SYSTEM

The custom-made imaging systems at the Crystallomics Core (*Figure 1.1c*) incorporate 768 plate hotels at each of two temperatures (4°C and 20°C). After setup with the Hydra Plus One, plates are transferred to the appropriate temperature storage and imaging hotel via mobile cassettes, which can be preloaded and inserted into the system. The position of the plate in the hotel is recorded via a Microsoft ACCESS database. Upon loading, plates are assigned an imaging schedule, with plates typically imaged at 3, 7, 14, and 28 days before being removed from the system. Schedules are generated in advance and consist of simple comma-delimited text files containing the plate name, barcode, and the day scheduled for imaging. Each morning the appropriate plates are imaged, with plates transferred from the hotel to the imager and vice versa via a robot arm (*Figure 1.1c*). Imaging is performed with a Veeco/Optimag Oasis 1700 equipped with a 5×-magnification objective fixed focus microscope. The system records 96 images per plate in a little over 90 seconds, and these are stored as 256 grey-level Microsoft bitmap images containing 1024 × 1024 pixels. Images are transferred automatically to an image server each night, after which the relevant information for each image is deposited in the database. After analysis with annotation software, the images are compressed to jpeg format, resulting in a reduction to 1/20 the space, from 1 Mb to 50 kb.

IMAGE ANALYSIS SOFTWARE

To date, we have collected over 6 million images from approximately 20 000, 96-well plates. For this number of experiments, crystal detection and annotation of crystallization experiments is important. Broad sparse matrix screening with multiple imaging time points generates a large number of images, analysis of which presents a considerable challenge. Unfortunately, the highly individual behaviour of each drop considerably complicates the automation of this task. An ideal system would automatically classify each crystallization experiment so that data can be collated on the reaction of the protein to a particular condition. With this approach, it may be possible to deduce subtle trends within crystallization conditions. This is not a trivial task and, although a number of software strategies are available, none as yet appear to have proven capable of analysing experiments with the required level of accuracy (Zuk and Ward, 1991; Spraggon *et al.*, 2002; Wilson, 2002; Cumbaa *et al.*, 2003; Bern *et al.*, 2004).

Within the Crystallomics Core, a custom annotation program, Crystal Experiment Evaluation Program (CEEP) (Spraggon *et al.*, 2002), has been developed that classifies and annotates the images in one of six categories: (0) experiment mistake; (1) clear drop; (2) homogeneous precipitate; (3) inhomogeneous precipitate; (4) crystal hit; (5) mountable crystal. The system works by clustering a number of features

together by means of a self-organizing neural net or Kohonen Net. The neural net was trained previously with a number of hand-annotated images. Twelve features in all can be used: six from edge detection methods, and six from texture-like methods. Using this methodology, the system can achieve a 70% classification accuracy, but with a large number of false positives (Spraggon *et al.*, 2002). Other use of texture-like features, such as those reported by Saitoh and colleagues, report a 90.6% agreement with human annotators using more textures and a linear discrimination classification scheme, suggesting that a reliable automatic system for classification may be on the horizon (Saitoh *et al.*, 2005). Until this time, annotation at the JCSG is also performed manually by visual inspection. Experiments are classified into three categories: unclassified; crystal hit; and mountable crystal (*Figure 1.2a*). Even without a contribution from automated drop analysis, the effect on productivity of scheduled automatic image collection, collation, and annotation should not be underestimated.

CRYSTAL OPTIMIZATION

Although reasonable arguments have been made that random screens are all that is necessary to grow diffracting crystals (Rupp, 2003), often a hit in a coarse screen needs to be optimized to produce crystals able to be mounted or merely to produce a large number of mountable crystals. To further speed and rationalize fine screening, a graphical user interface (GUI), NUDGE, has been developed to automatically generate fine screens from all of the coarse screens stored in the database (*Figure 1.2c*). The algorithm, written in Python and platform independent, identifies the chemical components required to produce a particular crystallization condition within the database, then, using this condition as the centroid of the screen, extends the condition in two dimensions, generally the pH and concentration of the major precipitant. If, after performing these calculations, the fine screen is unable to be made with the available stock solutions, then the ranges of the parameters are iteratively altered until the fine screen can be made. Having a default fine screen generator can lead to an efficient optimization procedure requiring little manual intervention. Manual mixing of the fine screen components can, however, be very labour intensive, particularly for high molecular weight polyethylene glycols (PEGs) where high viscosity is an issue. Fine screens are largely carried out within the GNF/JCSG core by the Alchemist system (RoboDesign, Inc., USA), a robotic mixing station capable of mixing any of the default fine screens to high precision.

Salvage pathways

Initially, a two-tiered approach is applied at the JCSG to screen for readily crystallizable proteins (Lesley *et al.*, 2002), with around 30% of targets expressing and crystallizing in the first round of experiments. However, many targets, particularly eukaryotic proteins, require a more significant investment of resources in order to obtain a structure. For many targets, it may be expected that expression in *Escherichia coli* will never be feasible due to the requirement for specialized eukaryotic post-translational processing, and for these there are many other well-known expression systems available.

Within the existing pipeline, methods are required to pursue targets that do not crystallize after initial screens of the native protein. Generally, there are three routes to success in attempting to obtain crystals after failed trials. One method is to change the conditions under which the protein is produced. Frequently, something as simple as changing the formulation of the proteins' final buffer or the purification protocols is enough, and it has been noted that some proteins that fail to crystallize after tier 1 screening will crystallize after further purification (Page *et al.*, 2003). Another route is to alter the protein, typically by truncation, mutation, or chemical modification. Finally, changes or extensions may be made to the crystallization screen, despite having sampled a large number of conditions in crystallization space. Crystallization condition space is so large and multivariate that it is impossible to cover it adequately, particularly when proteins are exquisitively sensitive to a specific condition, and it has been argued that further increasing the number of crystallization conditions screened provides little benefit (Segelke, 2001). However, within the existing GNF high-throughput pipeline, there are numerous techniques based upon the first two of these approaches that have been applied successfully to recalcitrant targets.

STABILITY, SOLUBILITY, DISPERSITY

For crystallization to occur, the target protein first must survive purification and be delivered to the droplet in solution at a relatively high concentration, and some proteins may drop out of the pipeline because of instability or insolubility in the normal pipeline buffers. Often, proteins need to be expressed and purified in the presence of cofactors to enhance solubility or stability, but this may be impossible if the cofactors are unknown. To enhance protein stability for proteins exhibiting high levels of aggregation, data mining crystal trial images for clear droplets can suggest improved buffer conditions (Page *et al.*, 2005). For the characterization of proteins in solution, further biophysical techniques are required. There are numerous reports of the correlation between monodispersity and successful crystallization. One study reported that of 66 proteins examined by dynamic light scattering (DLS), 77% with monodisperse distributions in dilute solution crystallized, whereas only 8% with a polydisperse distribution did so (D'Arcy, 1994). The osmotic second virial coefficient (B_{22}), as determined by static light scattering (SLS), has also been suggested to correlate with predisposition to crystallization (George and Wilson, 1994).

As well as providing a further biophysical technique for quality control, light scattering and size exclusion can provide important information on the true multimeric state of a sample for later analysis. The biological unit (BU) of a protein, which is the recognized or reputed functional macromolecule, can contain one, a portion, or multiple copies of the asymmetric unit (ASU). It can be predicted from the crystallographic data deposited in the PDB (Berman *et al.*, 2002) using the protein quaternary structure server (PQS) (Henrick and Thornton, 1998). Furthermore, functional or biochemical annotation can be used to predict the BU. Experimental methods are much preferred, however, as computational methods are not always reliable (McMullan *et al.*, 2005). The dispersity, molecular weight (MW), and oligomeric state of a given protein in the production and crystallization pipeline at GNF are determined using size exclusion chromatography (SEC) in combination

with SLS. An Agilent 1100 HPLC is used with a Superdex 200 column (GE Healthcare) in series with miniDAWN SLS and Optilab DSP differential refractive index detectors, both from Wyatt Technology, Inc. (USA).

USE OF ORTHOLOGUES

No *a priori* methods exist that can accurately predict protein solubility. It has been noted often that small changes in amino acid sequence can have dramatic consequences on solubility (Eberstadt *et al.*, 1998). Early in the evolution of protein crystallography, it was realized that when a protein from one species does not crystallize, the equivalent protein from a different species will (Kendrew *et al.*, 1954; Campbell *et al.*, 1972). For these reasons, it is a well-established practice to attempt expression and crystallization trials with a series of orthologues to increase the probability of success. Since the effect of a protein's variable regions on heterologous expression and crystallization are fortuitous, it does not necessarily follow that one orthologue will contain the best combination. It has been demonstrated that 'DNA shuffling' can generate crystallizable protein from donor orthologues that do not themselves crystallize (Keenan *et al.*, 2005). This approach is particularly applicable to structural genomics pipelines, since the constructs can be designed without use of prior knowledge, although efficacy remains to be proven in this setting.

SURFACE RESIDUE ENGINEERING

Another method shown to enhance crystallizability is to engineer point mutations within the protein of interest (Derewenda, 2004a,b). This procedure may aim to remove hydrophobic residues suspected of triggering nonspecific aggregation (Mateja *et al.*, 2002). For example, scanning mutagenesis targeting hydrophobic residues leads to identification of F185K mutant HIV integrase as suitable for structural studies (Jenkins *et al.*, 1995). A second approach is to remove flexible side chains, which are thought to inhibit crystallization by increasing surface conformational entropy. Lo Conte and colleagues discovered that although lysines and glutamates constitute 14.9 and 10.3% of surface residues in a panel of 75 structures, respectively, they occur at protein–protein complex interfaces only 5.4 and 4.9% of the time (Lo Conte *et al.*, 1999). These techniques have the disadvantage of requiring mutagenesis of the target gene, and also require an element of speculation as to which residues are located at the protein surface.

Suffering neither of these drawbacks, the technique of reductive methylation is becoming increasingly popular. This technique involves chemical modification of lysine side chains in the mature protein, such that only solvent-exposed side chains are altered, providing maximum effect on the crystallization surfaces with minimum damage to residues important in folding and stabilization. Originally used to determine the structure of chicken myosin subfragment 1 (Rayment *et al.*, 1993), reductive methylation is being adopted in structural genomics efforts to improve crystallization (Rypniewski *et al.*, 1993; Kobayashi *et al.*, 1999; Kurinov *et al.*, 2000; Schubot and Waugh, 2004), and has been shown to be critical for crystallization in some cases (Schubot and Waugh, 2004).

REMOVAL OF FLEXIBLE REGIONS

Current dogma suggests that inherent disorder within proteins can prevent crystallization by inhibiting the formation of stable crystal contacts and thereby reducing the probability of nucleation. One of the criteria in the design of a construct that has a good chance of crystallization is to minimize the inclusion of flexible regions. Computational approaches examining sequence complexity such as SEG (Wootton and Federhen, 1993) or secondary structure prediction algorithms, are commonly utilized, but an ideal system would incorporate experimental data into the system. Intrinsic flexibility is considered to interfere with crystal formation, and often the goal in truncating a protein is to remove regions that are not expected to form compact regular folds. Although many proteins with intact leader sequences have been crystallized, only in rare cases do the His tags show order within the crystal structure. It is an obvious first step, therefore, to remove the 12 residue leader and His tag, which can be performed after affinity purification by incorporation of a protease cleavage site after the tag. Partial proteolysis has often been applied to identify domain boundaries and remove flexible regions (Porter, 1973). Similarly, mass spectrometry can identify truncated species produced during purification through the action of endogenous host proteases to provide information for recloning, if necessary.

DEUTERIUM EXCHANGE MASS SPECTROMETRY (DXMS)

Biophysical measurements of flexibility and exposure to solvent can also be provided by deuterium exchange mass spectrometry (DXMS). The DXMS method (Woods, 2001a,b; Woods and Hamuro, 2001; Hamuro *et al.*, 2002a,b, 2003; Pantazatos *et al.*, 2004) provides an attractive alternative to computational approaches by coupling the labelling of flexible and solvent-exposed regions in the native protein with simple and sensitive detection and analysis.

Peptide amide hydrogen-exchange techniques have been used to study the thermodynamics of protein conformational change and the mechanisms of protein folding for more than forty years, and have recently proved effective in the investigation of domain structure, regional stability, and protein function (Englander *et al.*, 1997; Engen and Smith, 2001). The technique takes advantage of exchange rates between amide hydrogens and the surrounding solvent. As would be expected, one of the principal factors affecting the rates of exchange of the hydrogen with the solvent is the accessibility of the residue to the surrounding water.

Measurement of the local hydrogen exchange rate for continuous regions of proteins can then provide information about the dynamics of individual residues within the proteins. The rate of hydrogen/deuterium exchange in residues of folded proteins varies by up to a billion-fold between a buried amide hydrogen compared to one exposed to solvent. The DXMS method has become a standard salvage method for proteins that produce few or no crystallization hits in initial coarse screens. It rapidly and precisely identified regions of disorder using as little as 100 μg of protein, which can then be used to inform the design of child constructs containing selected deletions. The resulting constructs have shown a marked improvement in crystallization propensity (Pantazatos *et al.*, 2004; Spraggon *et al.*, 2004), and it is

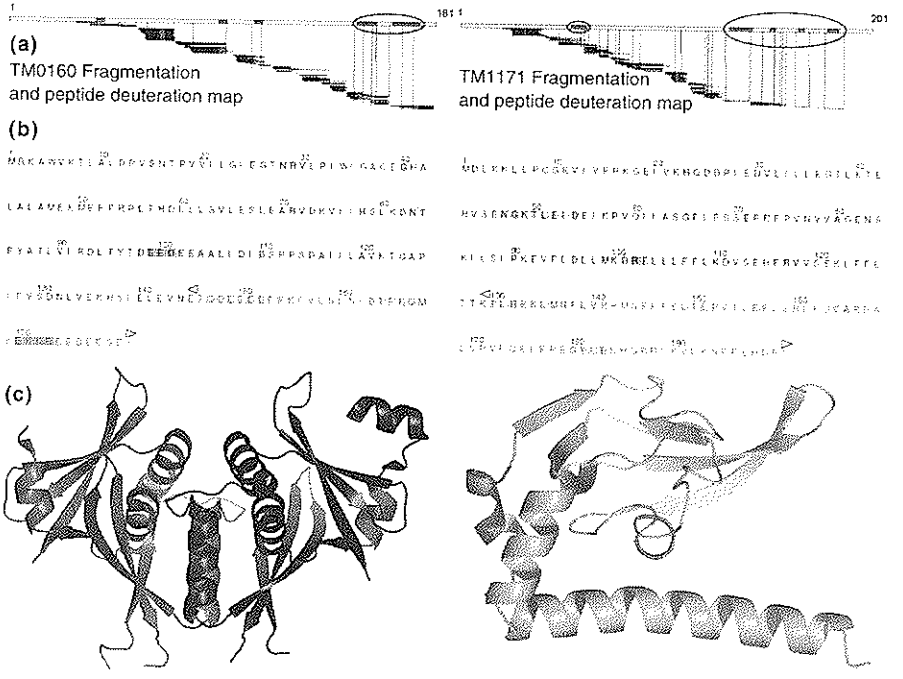

Figure 1.3. DXMS mapping of TM0160 (left) and TM1171 (right). (a) DXMS peptide analysis showing tryptic peptides mapped onto the gene schematic, with flexible regions shaded. (b) Protein sequences. Residues indicated to be flexible by DXMS are shaded, residues indicated to be flexible by SEG are shaded and boxed, residues indicated as flexible by both methods are shaded darkly, and boxed. Residues deleted as a consequence of the DXMS analysis are shown flanked by triangles. (c) Structures resulting from the modified constructs.

estimated that between 20 and 40% of failed proteins can be rescued by this method (Pantazatos *et al.*, 2004). This has been demonstrated by two examples from the genome of *T. maritima*, TM0160 and TM1171 (Spraggon *et al.*, 2004). These two proteins, which, in their full-length states, were resistant to crystallization attempts, were used to demonstrate the DXMS utility. In both cases, constructs were truncated at regions where the DXMS data indicated flexibility. The first, TM0160, is a novel fold and was truncated at its C-terminus to yield viable crystals and a structure (*Figures 1.3b* and *1.3c*). The second was TM1171, a transcriptional regulator protein that probably requires a DNA substrate to form a stable structure. The designed construct for TM1171 excised a sub-domain from the C-terminus as flexibility in this region probably inhibited crystallization (*Figures 1.3b* and *1.3c*). These results demonstrate that DXMS can provide a simple and rapid means to give meaningful data as to where to terminate/separate domains to provide more stable and ordered constructs, with no prior knowledge of the protein structure or function.

Novel phasing techniques and biotechnology

As for expression, purification, and crystallization, reliable and highly generic methods are required for experimental phasing in a high-throughput setting.

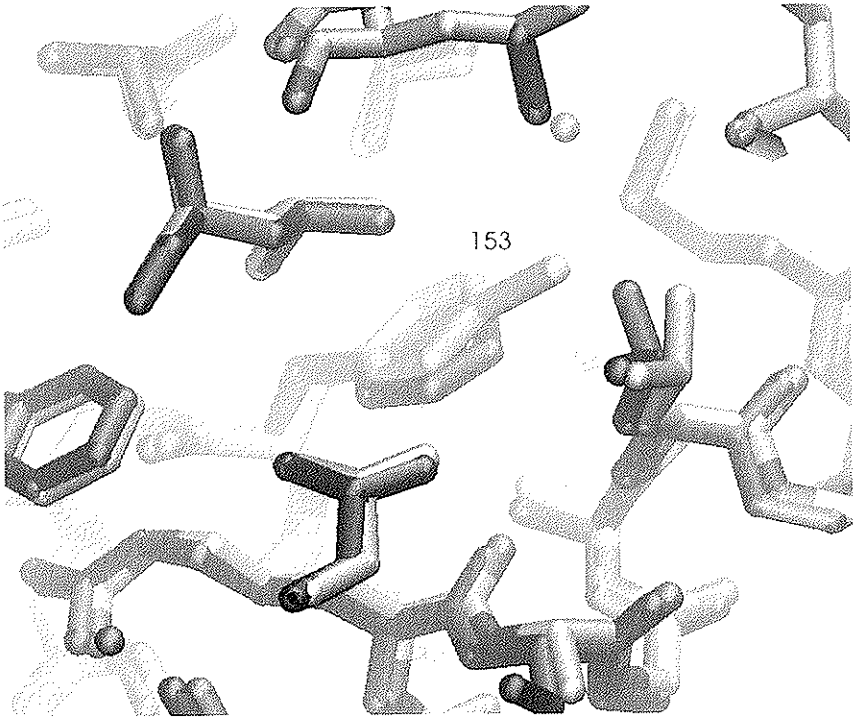

Figure 1.4. Structural alignment of native lysozyme (dark grey) with iodoPhe-153 substituted lysozyme (light grey). Residue 153 is labelled, with the iodine atom indicated by a black dot. This figure shows that even when introduced into the protein hydrophobic core, the iodoPhe residue is accommodated with only small conformational adjustments in the environment. Lysozyme incorporating iodoPhe was used to demonstrate the utility of experimental phasing using unnatural amino acids and in-house X-ray facilities (Xie *et al.*, 2004). Figure produced with visual molecular dynamics (VMD) (Humphrey *et al.*, 1996).

Selenomethionine incorporation fulfils these criteria very successfully, but requires access to tunable wavelength synchrotron beamlines. Furthermore, some targets appear resistant to expression or crystallization as selenomethionine derivatives, and some proteins simply lack a sufficient number of methionine residues for phasing in this manner to succeed. Site-specific incorporation of unnatural amino acids is finding utility in a number of fields, and one application is the incorporation of heavy/anomalous scattering atoms into protein. This technology, developed by Schultz and colleagues, is readily applicable to both prokaryotic and eukaryotic systems (Wang *et al.*, 2001; Chin *et al.*, 2003). The basis for this approach is the incorporation of iodoPhe (p-iodo-L-phenylalanine) into a target protein at an engineered amber stop codon. To facilitate this in *E. coli*, a previously generated *Methanococcus jannaschii* $\text{tRNA}^{\text{Tyr}}_{\text{CUA}}$–TyrRS pair was used, which is orthologous to the *E. coli* tRNA sythetases and shows no nonspecific cross-reactivity (Chin *et al.*, 2003).

Crystals of iodoPhe engineered T4 lysozyme were grown and diffracted on a conventional in-house rotating anode X-ray system. It was shown that detection of the iodine atom within the core of lysozyme and initial phasing of the data could be achieved with only twice as much data as would be collected normally for a native

crystal (Xie *et al.*, 2004). This is a considerable reduction in the data redundancy required for similar experiments, for example those using sulfur (Dauter *et al.*, 1999). *Figure 1.4* shows the resulting structure aligned with native protein, which demonstrates that the introduced iodine atom is accommodated with only minor conformational changes.

Conclusions

Built on previous advances in the field of structural biology, the past six years have seen a flurry of progress aimed at accelerating the structure determination process. This has produced a vast number of new and interesting protein structures into the public domain and is also providing an interface for the transition of macromolecular X-ray crystallography from a difficult and laborious science to a mainstream laboratory tool for the study of biological molecules. The exponential increase in the number of macromolecule structures in the PDB was predicted in the 1970s by Richard Dickerson. The Dickerson equation predicts the number of structures deposited in a year to increase exponentially from 1960. The prediction was within 57 structures in 2001 and continues to be accurate. Despite the fact that this pace is unsustainable, it is exciting that atomic-resolution structures representative of the majority of protein families could be available within the next twenty years. With heightened productivity, computational and robotic tools have been developed to streamline processes requiring human intervention and to keep abreast of the mountain of data produced. Finally, the application of new and old crystallization tools within high-throughput pipelines should ensure that the structures keep coming at increasing rates.

Acknowledgements

We would like to thank all the members of the JCSG for their tireless work within the project, Peter Schultz for continuing support and scientific expertise, and the GNF engineering staff for development and support of robotic systems. This work was supported by NIH Protein Structure Initiative grants P50-GM 62411 and U54 GM074898 from the National Institute of General Medical Sciences (www.nigms.nih.gov).

References

ABOLA, E., KUHN, P., EARNEST, T. AND STEVENS, R.C. (2000). Automation of X-ray crystallography. *Nature Structural Biology* **7** Suppl., 973–977.

BERMAN, H.M., BATTISTUZ, T., BHAT, T.N. *ET AL.* (2002). The Protein Data Bank. *Acta Crystallographica. Section D: Biological Crystallography* **58**, 899–907.

BERN, M., GOLDBERG, D., STEVENS, R.C. AND KUHN, P. (2004). Automatic classification of protein crystallization images using a curve tracking algorithm. *Journal of Applied Crystallography* **37**, 279–287.

BROWN, J., WALTER, T.S., CARTER, L., ABRESCIA, N.G.A. *ET AL.* (2003). A procedure for setting up high-throughput nanolitre crystallization experiments. II. Crystallization results. *Journal of Applied Crystallography* **36**, 315–318.

BURLEY, S.K. (2000). An overview of structural genomics. *Nature Structural Biology* **7** Suppl., 932–934.

CAMPBELL, J.W., DUEE, E., HODGSON, G. *ET AL.* (1972). X-ray diffraction studies on enzymes in the glycolytic pathway. *Cold Spring Harbor Laboratory. Symposia on Quantitative Biology* **36**, 165–170.

CANAVES, J.M., PAGE, R., WILSON, I.A. AND STEVENS, R.C. (2004). Protein biophysical properties that correlate with crystallization success in *Thermotoga maritima*: maximum clustering strategy for structural genomics. *Journal of Molecular Biology* **344**, 977–991.

CHIN, J.W., CROPP, T.A., ANDERSON, J.C., MUKHERJI, M., ZHANG, Z. AND SCHULTZ, P.G. (2003). An expanded eukaryotic genetic code. *Science* **301**, 964–967.

CUMBAA, C.A., LAURICELLA, A., FEHRMAN, N. *ET AL.* (2003). Automatic classification of sub-microlitre protein-crystallization trials in 1536-well plates. *Acta Crystallographica. Section D: Biological Crystallography* **59**, 1619–1627.

D'ARCY, A. (1994). Crystallizing proteins – a rational approach? *Acta Crystallographica. Section D: Biological Crystallography* **50**, 469–471.

DAUTER, Z., DAUTER, M., DE LA FORTELLE, E., BRICOGNE, G. AND SHELDRICK, G.M. (1999). Can anomalous signal of sulfur become a tool for solving protein crystal structures? *Journal of Molecular Biology* **289**, 83–92.

DEREWENDA, Z.S. (2004a). The use of recombinant methods and molecular engineering in protein crystallization. *Methods* **34**, 354–363.

DEREWENDA, Z.S. (2004b). Rational protein crystallization by mutational surface engineering. *Structure* **12**, 529–535.

DIDONATO, M., DEACON, A.M., KLOCK, H.E., MCMULLAN, D. AND LESLEY, S.A. (2004). A scaleable and integrated crystallization pipeline applied to mining the *Thermotoga maritima* proteome. *Journal of Structural and Functional Genomics* **5**, 133–146.

EBERSTADT, M., HUANG, B., CHEN, Z. *ET AL.* (1998). NMR structure and mutagenesis of the FADD (Mort1) death-effector domain. *Nature* **392**, 941–945.

ENGEN, J.R. AND SMITH, D.L. (2001). Investigating protein structure and dynamics by hydrogen exchange MS. *Analytical Chemistry* **73**, 256A–265A.

ENGLANDER, S.W., MAYNE, L., BAI, Y. AND SOSNICK, T.R. (1997). Hydrogen exchange: the modern legacy of Linderstrom–Lang. *Protein Science* **6**, 1101–1109.

GEORGE, A. AND WILSON, W.W. (1994). Predicting protein crystallization from a dilute solution property. *Acta Crystallographica. Section D: Biological Crystallography* **50**, 361–365.

GILLILAND, G.L., TUNG, M. AND LADNER, J. (1996). The Biological Macromolecule Crystallization Database and NASA Protein Crystal Growth Archive. *Journal of the Research of the National Institute of Standards and Technology* **101**, 309–320.

GREEN, D.W., INGRAM, V.M. AND PERUTZ, M.F. (1954). The structure of haemoglobin. IV. Sign determination by the isomorphous replacement method. *Royal Society of London. Proceedings. Series A* **225**, 287–307.

HAAS, C. AND DRENTH, J. (1999). Understanding protein crystallization on the basis of phase diagrams. *Journal of Crystal Growth* **196**, 388–394.

HAMURO, Y., BURNS, L.L., CANAVES, J.M., HOFFMAN, R.C., TAYLOR, S.S. AND WOODS, V.L., JR. (2002a). Domain organization of D-AKAP2 revealed by deuterium exchange-mass spectrometry (DXMS). *Journal of Molecular Biology* **4**, 703–714.

HAMURO, Y., WONG, L., SHAFFER, J. *ET AL.* (2002b). Phosphorylation-driven motion in the COOH-terminal Src kinase, Csk, revealed through enhanced hydrogen–deuterium exchange and mass spectrometry (DXMS). *Journal of Molecular Biology* **323**, 871–881.

HAMURO, Y., ZAWADZKI, K.M., KIM, J.S., STRANZ, D., TAYLOR, S.S. AND WOODS, V.L., JR. (2003). Dynamics of cAPK type IIb activation revealed by enhanced amide H/2H exchange mass spectrometry (DXMS). *Journal of Molecular Biology* **327**, 1065–1076.

HANSEN, C. AND QUAKE, S.R. (2003). Microfluidics in structural biology: smaller, faster...better. *Current Opinion in Structural Biology* **13**, 538–544.

HANSEN, C.L., SKORDALAKES, E., BERGER, J.M. AND QUAKE, S.R. (2002). A robust and scaleable microfluidic metering method that allows protein crystal growth by free interface diffusion. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 16531–16536.

HELLIWELL, J.R. (1998). Synchrotron radiation facilities. *Nature Structural Biology* **6**, 614–617.

HENRICK, K. AND THORNTON, J.M. (1998). PQS: a protein quaternary structure file server. *Trends in Biochemical Sciences* **23**, 358–361.

HENDRICKSON, W.A. (1991). Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51–58.

HENDRICKSON, W.A. (1999). Maturation of MAD phasing for the determination of macromolecular structures. *Journal of Synchrotron Radiation* **6**, 845–851.

HENDRICKSON, W.A., HORTON, J.R. AND LEMASTER, D.M. (1990). Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO Journal* **9**, 1665–1672.

HOPPE, W. (1957). Die faltmolekülmethode – eine neue methode zur bestimmung der kristallstruktur bei ganz oder teilweise bekannter molekülstruktur. *Acta Crystallographica* **10**, 750–751.

HOSFIELD, D., PALAN, J., HILGERS, M., SCHEIBE, D., MCREE, D.E. AND STEVENS, R.C. (2003). A fully integrated protein crystallization platform for small-molecule drug discovery. *Journal of Structural Biology* **142**, 207–217.

HUMPHREYS, W., DALKE, A. AND SCHULTEN, K. (1996). VMD – visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33–38.

JANCARIK, J. AND KIM, S.-H. (1991). Sparse matrix sampling: a screening method for crystallization of proteins. *Journal of Applied Crystallography* **24**, 409–411.

JENKINS, T.M., HICKMAN, A.B., DYDA, F., GHIRLANDO, R., DAVIES, D.R. AND CRAIGIE, R. (1995). Catalytic domain of human immunodeficiency virus type 1 integrase: identification of a soluble mutant by systematic replacement of hydrophobic residues. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 6057–6061.

KEENAN, R.J., SIEHL, D.L., GORTON, R. AND CASTLE, L.A. (2005). DNA shuffling as a tool for protein crystallization. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 8887–8892.

KENDREW, J.C., PARRISH, R.G., MARRACK, J.R. AND ORLANS, E.S. (1954). The species specificity of myoglobin. *Nature* **174**, 946–949.

KOBAYASHI, M., KUBOTA, M. AND MATSUURA, Y. (1999). Crystallization and improvement of crystal quality for X-ray diffraction of maltooligosyl trehalose synthase by reductive methylation of lysine residues. *Acta Crystallographica. Section D: Biological Crystallography* **55**, 931–933.

KRUPKA, H.I., RUPP, B., SEGELKE, B.W. *ET AL.* (2002). The high-speed hydra-plus-one system for automated high-throughput protein crystallography. *Acta Crystallograhica* **D58**, 1523–1526.

KURINOV, I.V., MAO, C., IRVIN, J.D. AND UCKUN, F.M. (2000). X-ray crystallographic analysis of pokeweed antiviral protein-II after reductive methylation of lysine residues. *Biochemical and Biophysical Research Communications* **275**, 549–552.

LESLEY, S.A. (2001). High-throughput proteomics: protein expression and purification in the postgenomic world. *Protein Expression and Purification* **22**, 159–164.

LESLEY, S.A., KUHN, P., GODZIK, A. *ET AL.* (2002). Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 11664–11669.

LICHTY, J.J., MALECKI, J.L., AGNEW, H.D., MICHELSON-HOROWITZ, D.J. AND TAN, S. (2005). Comparison of affinity tags for protein purification. *Protein Expression and Purification* **41**, 98–105.

LO CONTE, L., CHOTHIA, C. AND JANIN, J. (1999). The atomic structure of protein–protein recognition sites. *Journal of Molecular Biology* **285**, 2177–2198.

LUFT, J.R., RAK, D.M. AND DETITTA, G.T. (1999). Microbatch macromolecular crystallization in micropipettes. *Journal of Crystal Growth* **196**, 450–455.

LUFT, J.R., COLLINS, R.J., FEHRMAN, N.A., LAURICELLA, A.M., VEATCH, C.K. AND DETITTA, G.T. (2003). A deliberate approach to screening for initial crystallization conditions of biological macromolecules. *Journal of Structural Biology* **142**, 170–179.

MATEJA, A., DEVEDJIEV, Y., KROWARSCH, D. *ET AL.* (2002). The impact of Glu→Ala and Glu→Asp mutations on the crystallization properties of RhoGDI: the structure of RhoGDI at 1.3 A resolution. *Acta Crystallographica. Series D: Biological Crystallography* **58**, 1983–1991.

MCMULLAN, D., CANAVES, J.M., QUIJANO, K. *ET AL.* (2005). High-throughput protein production for X-ray crystallography and use of size exclusion chromatography to validate

or refute computational biological unit predictions. *Journal of Structural and Functional Genomics* **6**, 135–141.

MORRIS, R., LAMZIN, V.S. AND PERRAKIS, A. (1999). Protein microcrystals and the design of a microdiffractometer: current experience and plans at EMBL and ESRG/ID 13. *Nature Structural Biology* **6**, 458–463.

PAGE, R. AND STEVENS, R.C. (2004). Crystallization data mining in structural genomics: using positive and negative results to optimize protein crystallization screens. *Methods* **34**, 373–389.

PAGE, R., GRZECHNIK, S.K., CANAVES, J.M. *ET AL.* (2003). Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the *Thermotoga maritima* proteome. *Acta Crystallographica. Series D: Biological Crystallography* **59**, 1028–1037.

PAGE, R., PETI, W., WILSON, I.A., STEVENS, R.C. AND WUTHRICH, K. (2005). NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 1901–1905.

PANTAZATOS, D., KIM, J.S., KLOCK, H.E. *ET AL.* (2004). Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 751–756.

PORTER, R.R. (1973). Structural studies of immunoglobulins. *Science* **180**, 713–716.

RAYMENT, I., RYPNIEWSKI, W.R., SCHMIDT-BÄSE, K. *ET AL.* (1993). Three-dimensional structure of myosin subfragment-1: a molecular motor. *Science* **261**, 50–58.

ROSSMANN, M. AND BLOW, D. (1962). The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallographica* **A46**, 73–82.

RUPP, B. (2003). High-throughput crystallography at an affordable cost: The TB Structural Genomics Consortium Crystallization Facility. *Accounts of Chemical Research* **36**, 173–181.

RYPNIEWSKI, W.R., HOLDEN, H.M. AND RAYMENT, I. (1993). Structural consequences of reductive methylation of lysine residues in hen egg white lysozyme: an X-ray analysis at 1.8-A resolution. *Biochemistry* **32**, 9851–9858.

SAITOH, K., KAWABATA, K., ASAMA, H., MISHIMA, T., SUGAHARA, M. AND MIYANO, M. (2005). Evaluation of protein crystallization states based on texture information derived from greyscale images. *Acta Crystallographica* **D61**, 873–880.

SANTARSIERO, B.D., YEGIAN, D.T., LEE, C.C. *ET AL.* (2002). An approach to rapid protein crystallization using nanodroplets. *Journal of Applied Crystallography* **35**, 278–281.

SCHUBOT, F.D. AND WAUGH, D.S. (2004). A pivotal role for reductive methylation in the *de novo* crystallization of a ternary complex composed of *Yersinia pestis* virulence factors YopN, SycN and YscB. *Acta Crystallographica. Series D: Biological Crystallography* **60**, 1981–1986.

SEGELKE, B.W. (2001). Efficiency analysis of sampling protocols used in protein crystallization screening. *Journal of Crystal Growth* **232**, 553–562.

SPRAGGON, G., LESLEY, S.A., KREUSCH, A. AND PRIESTLE, J.P. (2002). Computational analysis of crystallization trials. *Acta Crystallographica. Series D: Biological Crystallography* **58**, 1915–1923.

SPRAGGON, G., PANTAZATOS, D., KLOCK, H.E., WILSON, I.A., WOODS, V.L., JR. AND LESLEY, S.A. (2004). On the use of DXMS to produce more crystallizable proteins – structures of the *Thermotoga maritima* proteins TM0160 and TM1171. *Protein Science* **13**, 3187–3199.

STEVENS, R.C. (2000). Design of high-throughput methods of protein production for structural biology. *Structure* **8**, R177–R185.

STEVENS, R.C. AND WILSON, I.A. (2001). Industrializing structural biology. *Science* **293**, 519–520.

TERWILLIGER, T.C. (2004). Structures and technology for biologists. *Nature Structural and Molecular Biology* **11**, 296–297.

WANG, L., BROCK, A., HERBERICH, B. AND SCHULTZ, P.G. (2001). Expanding the genetic code of *Escherichia coli*. *Science* **292**, 498–500.

WILSON, J. (2002). Towards the automated evaluation of crystallization trials. *Acta Crystallographica. Series D: Biological Crystallography* **58**, 1907–1914.

WOODS, V.L., JR. (2001a). Method for characterization of the fine structure of protein binding sites using amide hydrogen exchange. In: *US Patent # 6,331,400* USA.

WOODS, V.L., JR. (2001b). Methods for the high-resolution identification of solvent-accessible amide hydrogens in polypeptides or proteins and for the characterization of the fine structure of protein binding sites. In: *US Patent # 6,291,189* USA.

WOODS, V.L., JR. AND HAMURO, Y. (2001). High resolution, high-throughput amide deuterium exchange-mass spectrometry (DXMS) determination of protein binding site structure and dynamics: utility in pharmaceutical design. *Journal of Cellular Biochemistry* **37**, 89–98.

WOOTTON, J.C. AND FEDERHEN, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers and Chemistry* **17**, 149–163.

XIE, J., WANG, L., WU, N., BROCK, A., SPRAGGON, G. AND SCHULTZ, P.G. (2004). The site-specific incorporation of p-iodo-L-phenylalanine into proteins for structure determination. *Nature Biotechnology* **22**, 1297–1301.

ZUK, W.M. AND WARD, K.B. (1991). Methods of analysis of protein crystal images. *Journal of Crystal Growth* **110**, 148–155.