High-throughput crystallization and structure determination in drug discovery

Lance Stewart, Robin Clark and Craig Behnke

High-throughput protein X-ray crystallography offers an unprecedented opportunity to facilitate drug discovery. The key bottlenecks in the path from target gene to three-dimensional protein structure determination are defined. Special emphasis is placed on the concept that drug discovery projects are typically directed at a key protein target whose structure must be solved within a reasonable time frame to have an impact on the drug discovery process. The time-sensitive nature of structural data has placed growing pressure on the need to automate all aspects of protein crystallography, from gene identification to model building and refinement. Current technological innovations and strategies for automation are discussed with respect to the bottleneck they are intended to eliminate.

***Lance Stewart, Robin Clark and Craig Behnke**
Emerald BioStructures
(a subsidiary of MediChem
Life Sciences)
7869 N.E. Day Rd W.
Bainbridge Island
WA 98110, USA
*tel: +1 206 780 8535
fax: +1 206 780 8549
e-mail: lstewart@
emeraldbiostructures.com

▼ The explosion of information flowing from functional genomics research has led to the identification of an unprecedented number of potential therapeutic protein targets [1,2]. Consequently, drug discovery companies are faced with the daunting task of having to choose and prioritize the protein targets that are selected for drug discovery. This is typically achieved through target validation efforts that define the functional role(s) of the target under scrutiny. In the case of genes that encode proteins with unknown structure or function, there is the possibility that a rapid X-ray crystal structure determination could enable the identification of protein function through comparative structural analyses. This opportunity has spawned the field of 'structural genomics', which strives to solve at least one protein structure from all protein sequence families [3,4]. Funding from both public and private sources has fueled the establishment of at least 15 different multi-investigator projects in structural genomics [5–8].

Protein crystallography in drug design

The structural genomics efforts are focused almost entirely on soluble proteins of unknown structure or function [9] (http://www.nigms.nih.gov/news/meetings/structural_genomics_targets.html). By contrast, most drug discovery programs are directed at a specific protein target of known function, which is often a membrane bound protein. It is estimated that >50% of all major drug targets are membrane proteins [10,11]. As such, the protein targets that are chosen by structural genomics researchers are not necessarily of immediate value to drug discovery programs [12]. However, as the structural genomics projects have gotten underway, heavy emphasis has been placed on the development of automated high-throughput systems for protein expression, purification, crystallization and X-ray structure determination. The development of automated crystallography systems is poised to have an immediate and significant impact on the pharmaceutical industry by enabling protein–ligand co-crystal structures to be solved with increasing speed.

The demand for protein crystallography in drug discovery is driven by the need to understand exactly how small molecules (drug candidates) bind to their protein target. This information enables researchers to conduct medicinal chemistry projects in a more rational manner, taking advantage of structural information and applying it in structure-based approaches to making new compounds. However, structure-based drug design is complicated by the fact that computational screening methods often fail to accurately predict ligand-binding modes to protein targets [13–16], and the binding of a ligand to its

1359-6446/02/$ – see front matter ©2002 Elsevier Science Ltd. All rights reserved. PII: 1359-6446(02)02121-3

Box 1. Private companies that are engaged in high-throughput protein crystallography

3-Dimensional Pharmaceuticals, Exton, PA, USA
Abbott Laboratories, Chicago, IL, USA
Astex Technology, Cambridge, UK
Emerald BioStructures, Bainbridge Island, WA, USA
Integrative Proteomics, Toronto, Canada
Structural GenomiX, San Diego, CA, USA
Syrrx, San Diego, CA, USA
Vertex Pharmaceuticals, Cambridge, MA, USA

target can often result in large changes in protein conformation. Given the inadequacies of computational tools for predicting ligand-binding modes, there is a growing need for the crystal structure-determination of large numbers of ligand–protein complexes. To meet this demand within time scales that are reasonable for drug development programs, it is necessary that the process of crystal structure determination be automated and industrialized. Box 1 lists the private companies that are engaged in high-throughput protein X-ray crystallography.

The crystallography game

It is generally agreed that there are five key milestone events in a protein crystal structure-determination program: (1) the generation of open reading frame constructs that yield decent quantities of protein in one or another recombinant protein expression system; (2) the establishment of an effective purification protocol for the target protein and generation of sufficient pure protein to initiate crystallization trials; (3) crystallization, screening and optimization to produce X-ray diffraction quality crystals; (4) collection of sufficient diffraction data to obtain a quality electron density map; and finally (5) model building and refinement. The interface between each of these five steps can be a rate-limiting bottleneck in the entire process. If each of these steps were placed on the squares around a Monopoly board, there would be a number of squares in between that said ‘Go Back’ to some previous step. It is rare that a crystal structure is solved without landing on at least one ‘Go Back’ square.

Drug discovery programs demand that crystal structures of specific protein–ligand targets are determined in a reasonable time (6–12 months). This requires the use of a robust parallel approach to establish numerous alternative (back up) plans for leap-frogging potential dead ends that might be encountered during a structure-determination program. The parallel approach is more efficient at ‘leaving no rock unturned’ in a structure-determination program. It serves to increase the chance of a successful structure determination and, at the same time, reduces the chance of prematurely terminating a potentially successful structure-determination program. This is important because the lost opportunity cost for halting work on a structure-determination program could be equal to, or greater than, the cost savings afforded by the timely availability of atomic coordinates of a protein–ligand complex.

A highly parallel approach to crystal structure-determination requires that an automated system be capable of producing a large number of expression vectors, testing multiple expression systems, establishing and implementing 1–10 mg protein purification protocols, setting up iterative crystallization experiments, rapidly testing crystals for diffraction quality, collecting X-ray diffraction data, determining phases and automatically building and refining the final structure. Such a system does not exist today. However, the pieces are rapidly coming together and it is probable that it will become available within the next three to five years.

The database requirement

A very important consideration for automated and paralleled crystallization research is that the data generated at each step in the process need to be archived in an intelligent manner so that it can be used to efficiently optimize crystal growth in an iterative experimental approach [17]. Similarly, for high-throughput crystallography robotics hardware to be used effectively in drug development, it is imperative that the hardware is integrated with a relational database. In this way, a single platform can capture all experimental set-up parameters, prepare experimental screening plates, collect observation data, and allow researchers to design the next round of experiments.

The Biological Macromolecular Crystallization Database (BMCD) [18] represents a significant resource of information on crystal growth conditions. However, the BMCD only contains data on crystallizations that actually produced crystal structures. This is only a small fraction of the total number of crystallization trials that might have been performed to produce the final crystallization condition. To pursue protein crystal growth efficiently and intelligently, all crystallization data should be collected. Moreover, the captured data needs to be of sufficient detail to retain enough relational information that it can be interfaced with software that enables researchers to design follow-up experiments and drive robotic equipment to set up the experiments. The experimental data that are collected also need to be captured with a high level of detail so that artificial intelligence systems can search for correlative relationships in the data, and possibly even suggest the next round of crystal growth optimization experiments [19]. With these requirements in mind, several academic and industrial groups have created database systems for capturing and analyzing protein crystal growth data [19–23]. These systems include the Parallel Experiment Planning (PEP) system developed at the University of Pittsburgh (PA, USA) [19,22], CrystaLEAD™ [21] of Abbott Laboratories (Chicago, IL, USA), and Crystal Monitor™ of Emerald BioStructures (Bainbridge Island, WA, USA) [20].

Importantly, this database software approach to the analysis of crystallization trial data has begun to suggest that certain structural classes of proteins might crystallize in particular regions of an otherwise expansive crystallization space [22,24]. Furthermore, there are early hints from some of the structural-genomics program databases that there could be a significant correlation between amino acid content, protein solubility and crystallizability [25]. Thus, data mining of captured information has the power to accelerate the throughput of crystal structure determination significantly. The identification of complex correlations in experimental data is only possible if the data is collected in a consistent database format. Having a solid database approach to automation and experimentation must be emphasized as one of the underlying keys to a parallel approach in high-throughput crystallography for drug discovery.

Cloning and expression testing

The genomics revolution has provided an almost complete set of gene sequences from the human genome and several other species for target selection. The first, and possibly the most challenging, step in solving the structures of the chosen gene products is to produce the proteins encoded by these genes in a pure form and in sufficient quantities (1–10 mg).

There are several key criteria to be addressed when forming a strategy for protein production: source of DNA, appropriate transcription–translation system, factors affecting protein folding, efficient vector construction technology, affinity tags, protein modifications, purification systems and quality assurance.

Genomic DNA is an appropriate source for prokaryotic genes, but for human and higher eukaryote genes, introns interfere with the use of genomic DNA in most expression systems. Therefore, cDNA clones are the preferred source of protein-coding DNA for most expression efforts. There are several potential problems that arise when depending on the use of cDNA clones. Frequently, the clones are not full-length, thus requiring extensive recloning or patching with synthetic DNA derived from the genomic sequence. Public and private efforts are underway to construct complete full-length cDNA libraries for protein expression [26]. Available sequences for cDNA libraries, such as dbEST (the database of expressed sequence tags) [27] are often of low quality or misidentified, or the clones might encode undesirable splice-variants. The codon usage of a cDNA could be sub-optimal for the chosen expression system, especially for prokaryotic systems. Using synthetic gene construction could circumvent many of these problems. Whole gene synthesis involves the preparation of a set of oligonucleotides that, when assembled properly, will encode the desired gene sequence. Assembly of the oligonucleotides into a contiguous gene can be performed through sequential enzymatic ligation of overlapping oligonucleotides, through sequential PCR amplification of overlapping oligonucleotides, or by a combination of the two methods.

Box 2. Companies that are offering whole gene synthesis research services

Aptagen/Gene Forge, Herndon, VA, USA
Blue Heron Biotechnology, Bothell, WA, USA
Egea Biosciences, San Diego, CA, USA
Entelechon, Regensburg, Germany
Geneart, Regensburg, Germany
Integrated DNA Technologies, Coralville, IA, USA
Midland Certified Reagents, Midland, TX, USA
Sigma/Genosys, Sydney, Australia
Sloning, Munich, Germany
Qiagen/Operon, Alameda, CA, USA

Several considerations must go into the design of expression constructs for crystallography, including the location and nature of affinity tags for purification, site-specific proteolytic cleavage sites for tag removal, localization or secretion signals, mutations to alter secondary modifications, deletions to eliminate domains or sequences that promote degradation [28], or changes at the nucleotide level to alter codon usage and generally enhance transcription or translation. These considerations have led to the increasing demand for whole gene synthesis research services. Box 2 indicates the companies offering whole gene synthesis services. Although still quite expensive (US$ 5–20 per base pair), whole gene synthesis technology will continue to improve and will have an increasing impact on high-throughput crystallography and proteomics in general.

The choice of a protein expression system has important implications for throughput, cost-effectiveness and the probability of success in expressing many classes of proteins. The obvious choice for high throughput, easy set-up and cost-effective production is one of the many inducible bacterial expression systems, most of which are commercially available. However, most human proteins cannot be expressed in soluble form in bacteria [29] so a comprehensive effort must use additional expression systems.

Mammalian systems are capable of producing almost any human protein in the correctly folded and modified form but expression levels are usually low and mammalian expression is the least cost-effective type of system because of the complexity of mammalian cell culture. An effective alternative is the recombinant baculovirus–insect cell system. Most soluble human proteins can be expressed in soluble form and most mammalian posttranslational protein modifications are performed in insect cells [30]. This system enables multiple viral constructs encoding complimentary proteins to be coexpressed in the same cells to produce protein complexes or specific posttranslational modifications. Unfortunately, the expression levels in the baculovirus system are highly variable, and the set-up and maintenance of this system is less amenable to cost-effective high-throughput expression. Some progress has been made in modifying this system for high-throughput protein production [31]. Intermediate between insect and bacterial systems are the yeast systems. Originally *Saccharomyces cerevisiae* was the most commonly used yeast system. More recently, *Pichia pastoris* has emerged as the preferred yeast system because of its strong, highly inducible promoter system, stable genomic integration, posttranslational modifications that are more similar to the mammalian system, and a low level of background protein secretion [32]. Recently, *in vitro* transcription–translation has emerged as an alternative to host-vector systems for protein expression. When combined with continuous flow substrate exchange, such systems have shown promise in producing sufficient quantities of protein (0.5 mg of recombinant protein per ml of cell-free extract) for crystallography [33,34]. Regardless of the expression system chosen, an efficient method for generating large numbers of constructs must be considered. Fortunately, many automatable cloning systems have become available in recent years to replace the cumbersome restriction-enzyme-based cloning methods.

Protein purification and quality assurance

Once an acceptable level of expression has been achieved, there is the problem of purification. Because the lack of crystal growth could be the consequence of either the nature of the protein construct or its level of purity, it is desirable to make the effort to obtain highly purified and homogeneous protein to use in crystallization trials. If the trials fail to produce diffraction quality crystals, further scrutiny can then be placed on the nature of the construct. Effective chromatography protocols vary widely from one protein to the next, and even minor sequence alterations can have large effects on chromatographic behavior, which makes traditional methods impractical for parallel purification. Thus, it is necessary to incorporate affinity tags into the protein so that one (or a few) protocol(s) can be used. Affinity tags are also very useful in assessing and monitoring protein expression. Large tags, such as the maltose-binding protein, glutathione-*S*-transferase, thioreductase and the chitin-binding protein, could produce fusion proteins that are too large for high-level expression in some systems but could improve correct folding and stability. Small tags, such as the $(His)_6$ and various epitope tags [35], are easier to incorporate into expression constructs, especially where multiple tags are desirable. A protease cleavage site for tag removal should be considered, especially for the larger tags. Although proteases such as thrombin and enterokinase are commonly used for this purpose, the specificity of the Tobacco Etch Virus (TEV) protease appears superior to any other protease [36].

In cases where proteins are expressed as insoluble inclusion bodies, it is often desirable to purify the inclusion bodies by repeated sonification and washing. The purified insoluble protein can often be denatured in 6 M guanidine HCl or 8 M urea and then diluted into a ‘refolding buffer’ to produce refolded protein that is suitable for crystallization trials [37]. In an effort to streamline the search for an appropriate refolding buffer, we have developed a microscale refolding screen, wherein a small amount of denatured protein is mixed with an equal amount of a refolding solution (of which there are thousands) and then the refolding drop is allowed to equilibrate by vapor diffusion against a much larger volume of refolding solution. This experiment is performed in nearly the same manner as a typical crystallization experiment but in this case the desirable result is to see a large diluted ‘refolding’ drop that remains clear, which is an indication that refolding has occurred and the protein remains soluble. Refolding can be used in conjunction with metal-chelate chromatography, whereby a His-tagged protein is loaded onto a metal chelate matrix under denaturing conditions, the denaturant being removed by a series of equilibrations into the desired refolding buffer. The refolded protein can then be eluted from the column using an imidazole gradient.

Typically, crystallographic projects demand protein of high purity and homogeneity. Hence, final purified protein samples are subjected to a variety of biophysical characterizations, including SDS-PAGE, circular dichroism, isoelectric focusing, MS and dynamic light scattering [38]. All of these methods are designed to verify the nature and homogeneity of the protein sample. Although it is possible to crystallize proteins from impure mixtures (<85% pure), it is far more desirable to initiate crystallization trials with a pure and uniform protein sample (2–20 mg total, at ~2–10 mg ml^{-1} concentration).

Crystallization

Initial protein crystallization trial experiments routinely involve hanging-drop and sitting-drop vapor-diffusion methods [39]. The vapor-diffusion technique uses the evaporation and diffusion of volatile species (including water) between solutions of different concentrations as a means of achieving supersaturation. The sitting-drop vapor-diffusion experiment has advantages over the hanging-drop experiments, for the purposes of robotic image capture and analysis, because all crystallization drops are in a precision molded plasticware device and are, therefore, all in the same focal plane.

One other method for crystallizing proteins is known as the microbatch method [40]. A major distinction between the microbatch method and vapor-diffusion methods is that the microbatch trial is typically initiated with crystallization solutions with higher concentrations of precipitating agent, and tends to promote crystallization or precipitation on shorter time scales than vapor-diffusion methods, which are set up to allow a slow equilibration. Vapor-diffusion drops eventually reach equilibrium and the microbatch drops are allowed to continually lose vapor (slowly or quickly, depending on the nature of the oil overlay). Both techniques are amenable to high-speed automation.

Membrane proteins

In contrast to soluble proteins, only a handful of integral membrane protein structures have been determined. Membrane proteins tend to aggregate amorphously, instead of assembling into (three dimensional) 3D crystalline lattices [41]. To prevent such aggregation, solubilizing detergents are added to membrane protein samples. In general, the methods used to crystallize soluble proteins are not successful in crystallizing membrane proteins. Increasing the ionic strength for crystallizing membrane proteins, as is often done with soluble proteins, can cause the detergent to partition into a separate phase. When this occurs, the protein rapidly migrates into the detergent enriched phase, where depletion of the protein's solvation water results in its rapid denaturation [42]. This problem can be overcome by employing amphiphilic additives to prevent detergent phase separation. This enables membrane proteins to be crystallized in a manner similar to globular proteins [42]. Still, membrane protein crystallization successes have not been routine.

A novel technique for membrane protein crystallization is the lipidic cubic phase (LCP) method, developed by Landau and Rosenbusch [43,44] and used in the determination of the high resolution X-ray crystal structure of bacteriorhodopsin [44,45]. In the LCP method, the protein sample is mixed with lipid. Under the right conditions, the lipid forms a 3D cubic lattice with membrane protein embedded in the bilayer (Fig. 1a). Although the precise mechanism of protein crystallization by LCP is not yet fully understood, it is probable that the LCP lattice provides more opportunity for assembly into an ordered, 3D crystalline lattice (Fig. 1b). Integral membrane proteins crystallized using the LCP method include *Halobacterium salinarum* bacteriorhodopsin [44,45], *H. salinarum* halorhodopsin [46], *H. salinarum* sensory rhodopsin II [47], *Rhodobacter sphaeroides* photosynthetic reaction center [48], *Rh. viridis* photosynthetic reaction center [49] and *Rh. acidophila* light harvesting complex [50].

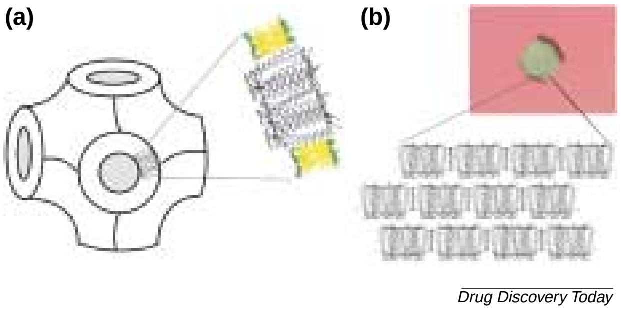


Figure 1. Lipidic cubic phase materials for membrane protein crystallization. **(a)** When mixed in the correct proportion, the solubilized membrane protein and lipid spontaneously form a cubic phase with the protein embedded in the bilayer. **(b)** This arrangement is conducive to the growth of tightly packed three-dimensional crystals comprised of stacked sheets of membrane proteins [43].

Automation

In the wake of the Human Genome Project, the field of structural genomics has placed increasing pressure on crystallographers to establish high-throughput automated systems for protein crystallization and crystal-structure determination. In general, the number of experiments that are necessary to determine the optimal crystallization conditions is large and, often, only a small amount of protein is available. Recently, crystallization robots have been developed to automate and speed up the experimental process. The robotics developed at the Hauptman-Woodward Medical Research Institute (Buffalo, NY, USA) now have the capacity to prepare and evaluate 40,000 crystallization experiments per day in microbatch mode [23]. The experiments are automated using robots with syringes to dispense the solutions and protein, and a robotic digital camera to record images of the crystallization experiments.

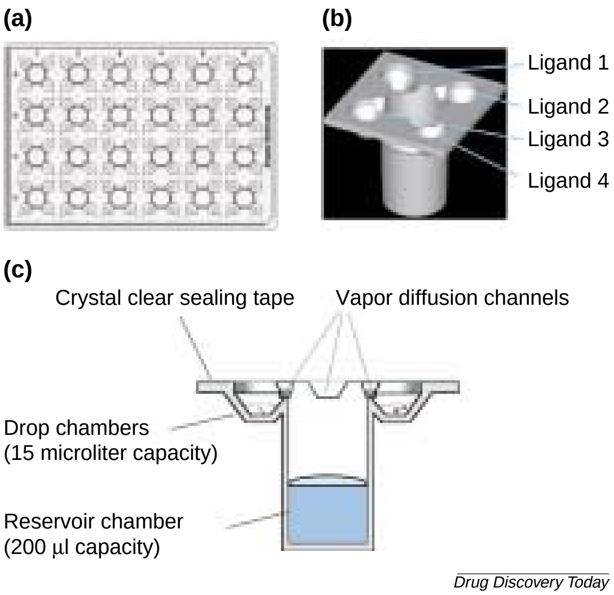


Figure 2. CombiClover™ plate. **(a)** Top view of the CombiClover plate with 24 separate clover crystallization chambers. **(b)** Tilt-angle view of a CombiClover crystallization chamber with lines to indicate the location of four different ligands that might be included in crystallizations. **(c)** A central reservoir for the crystallization solution is surrounded by four satellite chambers for the crystallization drops, enabling the simultaneous setup of four different protein samples. (US Patent 6,039,804) [55].

The robotic setup increases the number of initial experiments for each protein from the standard 96 conditions to 1536 conditions in a single plate.

Engineers from the Genomics Institute of the Novartis Research Foundation (San Diego, CA, USA) and crystallographers at Syrrx (San Diego, CA, USA) have collaborated to develop a crystallization robot and digital image capture system (Agincourt) that is capable of setting up and observing 60,000 micro-vapor diffusion crystallization experiments per day [51–53]. The volumes of the crystallization drops are in the 80 nl range, which greatly reduces the amount of protein required for screening, compared with conventional crystallization robots, such as the C-200 robot offered by Gilson-Cyberlab (Middleton, WI, USA), which requires 1–2 µl per experiment. The size of crystals obtained from these micro-drops can sometimes be of sufficient size to obtain high quality X-ray diffraction data.

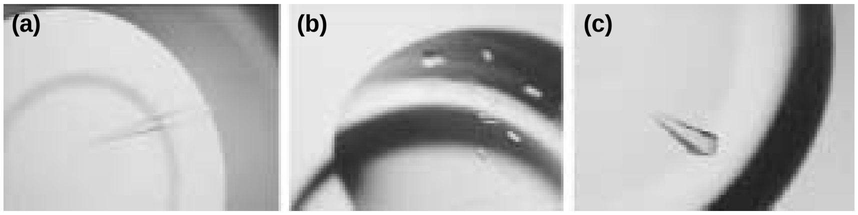
Automation specialists and software developers at Emerald BioStructures have generated robotic liquid-handling devices for mixing crystallization reagents from stock solutions (Matrix Maker™) and for setting up protein–ligand co-crystallization drops in 500 nl sitting-drop vapor-diffusion mode (Drop Maker™) using a patented plasticware device (CombiClover™) that connects four independent sitting-drop chambers to a common reservoir through vapor-diffusion channels (Fig. 2) [54,55]. The robots are capable of setting up 10,000 different protein–ligand crystallization drops in an 8 hr period (~105 plates with 96 drops per plate). The resulting plates are interrogated with a robotic stereomicroscope (Crystal Monitor Workstation™) that can capture a digital image for each of 96 crystallization drops on a plate in less than five minutes [54].

Each of the robotic crystallization systems described previously strive to gather protein crystallization data with high speed and the judicious use of protein sample. Like most screening applications, crystal growth is an iterative process that often requires optimization of several parameters. For example, a typical crystallization drop might contain crystallization solution (buffer, precipitating agent and salt), ligand, protein, cofactor, additive, reducing agent, and so on, each of which might require optimization in their nature, volume or concentration. Thus, crystallization experiments involve more complicated liquid-handling methods than typical enzyme assay screens. In addition, crystal growth assays do not have a single timed end-point but, rather, must be observed at multiple time-points. Given these factors, together with the growing demand in the drug industry for protein–ligand cocrystal structures, it is anticipated that crystallization robotics will continue to be improved by incorporating technologies, such as image analysis, to automatically detect crystal growth, and microfluidics to enable more efficient and complicated use of reagents.

Ligand-complex crystallization screening and crystal cracking assays

Screening for protein-crystal growth conditions that depend on the presence or absence of a ligand will increasingly become an important approach as crystallography is applied to drug discovery. As noted above, computationally derived models for ligand binding need to be confirmed with actual crystal structures. Therefore, when structure-based methods have been used to design the next generation of ligands, these too must be examined in high-resolution protein–ligand crystal structures to ensure that the design process is progressing as anticipated.

Crystal forms that only grow in the absence of a ligand (apocrystal forms) serve two possible functions. Some apocrystal forms can be soaked with multiple ligands and yield diffraction quality ligand-bound protein crystals [56]. In this case, multiple ligand-bound crystal structures can be derived rapidly from a single apocrystal form. Alternatively, many apocrystal forms will visibly crack upon being soaked with a ligand. This apocrystal cracking phenomenon is generally attributed to a ligand-induced conformational change that is incompatible with the packing arrangement within the crystal. This has been verified in a number of cases where the ligand–protein complex crystal structures have been solved and it is clear that significant conformational changes in the protein must have occurred upon ligand binding [57,58]. Thus, the crystal cracking phenomenon can actually be used as a screening tool for libraries of compounds that are tested by soaking apocrystals with individual ligands or mixtures of ligands. Visible crystal cracking is evidence of the presence of a ligand that can bind to the protein.



Drug Discovery Today

Figure 3. Drug-dependent crystal growth. **(a)** Crystals of a binary complex of human topoisomerase I bound covalently to duplex DNA substrate will not grow in the presence of topoisomerase I poisons and will crack if soaked in mother liquor solutions containing such compounds. **(b)** and **(c)** Crystals of a different morphology are produced under alternative crystal growth conditions, which require the presence of the anti-cancer drug Topotecan **(b)** or an indenoisoquinoline **(c)**. The ternary complex crystals contain drug that is bound to human topoisomerase I in covalent complex with duplex DNA [59]. Figs 3(a) and (c) are 30 × magnification; Fig. 3(b) is 10 × magnification.

Crystal forms that only grow in the presence of a ligand (ligand-bound crystal forms) can also be used as a screening tool for individual ligands or mixtures of ligands, wherein crystallization trials are performed under conditions that only produce crystals if a ligand is present. For example, we have used human DNA topoisomerase I (topo I) to identify crystal growth conditions that only produce crystals of a topo I–DNA complex when the crystallization drop contains a camptothecin type anti-cancer compound. Camptothecins bind to the topo I-DNA complex by intercalating into a single strand nick at the site of topo I cleavage [59]. The conditions used to obtain crystals of the ternary topo I-DNA–drug complex have proved useful for identifying additional crystal forms that only grow in the presence of intercalative compounds (Fig. 3). It seems probable that a similar approach for other protein–ligand complexes will serve both as the assay for ligand binding to protein, as well as the crystalline source for a crystal-structure determination, which would define exactly how the ligand in question is binding to the protein.

Data collection

As an increasing number of proteins are being crystallized, the demand for rapid X-ray data collection is also on the rise. There are several advantages of using synchrotron sources instead of in-house X-ray sources. Primarily, the intensity of the X-rays generated at synchrotrons is significantly greater than in-house sources, reducing the time required for data collection and increasing the diffraction limits of the crystals. Higher resolution diffraction data result in more detailed information about the structure of the protein.

In addition to synchrotrons, automation at the level of crystal mounting and data collection software has also contributed significantly towards rapid and improved diffraction data collection. The Automated Crystal Transport, Orientation and Retrieval (ACTOR) system developed by Abbott Laboratories (Abbott Park, IL, USA) is one such device [21,60]. ACTOR performs many functions that crystallographers would otherwise need to do by hand and it maximizes X-ray source usage by greatly reducing the time required for changing crystals between dataset collection. The X-ray system can run continuously with minimum operator intervention and interruption. The 'BLU-ICE' data collection and controlling software developed by researchers at the Stanford Synchrotron Research Laboratory (Palo Alto, CA, USA) (http://smb.slac.stanford.edu/public/bluice/overview. html) is a graphical front-end of an integrated motion control and data acquisition system for crystallographic data collection at synchrotron light sources. BLU-ICE serves as a model for the remote control of crystallographic beam lines. Together, automated sample mounting and remote data collection software at synchrotrons make a powerful tool for high-throughput data collection.

Structure determination

Two methods are in use for obtaining the phase information for structure determination. First, if the given protein shares at least 35% sequence identity with any protein whose 3D structure is available then the initial attempts to solve the structure will be by molecular replacement (MR). If MR approaches prove to be unsuccessful, then the alternate method would involve heavy atom phasing procedures. In the MR method, the homologous probe structure is fit to the experimental data using three rotational parameters and three translational parameters. The most popular programs that are available to perform MR searches are EPMR, AMoRe, and CNX (or CNS) [61–63]. EPMR uses a six-dimensional evolutionary search algorithm, whereas AMoRe and CNX use sequential rotation and translation searches.

As increasing numbers of new protein structures are provided by the field of structural genomics, it has been anticipated that MR methods will ultimately become the standard method for protein-crystal-structure determination. With this possibility in mind, several groups have automated the process by packaging the popular molecular replacement algorithms (AMoRe or EPMR) into a database application that enables a user to select run parameters and any number of possible search models [e.g. protein structures from the Protein Data Bank (PDB)] [52,64,65]. Portions of the application code can be paralleled for distribution over a network of computers. This reduces the time required for structure determination by a factor proportional to the number of machines in the network. In the cases where multiple search models give similarly high MR correlations, sequence comparisons can be very useful in eliminating false-positive results.

The other method used for obtaining the phasing information is by heavy atom phasing procedures. In all such procedures, the signal for obtaining phases arises from naturally occurring or artificially introduced heavy atoms, which cause changes in the intensities of the diffracted X-rays. Few native proteins contain heavy atoms, so most protein crystals must be soaked or cocrystallized with heavy atom compounds to produce a crystal that can be used for phasing. Although it might be easy to screen even hundreds of different heavy-atom soaking or cocrystallization conditions, heavy atom-derivatized crystals are often non-isomorphous with the native crystals, which can present computational problems in the structure determination. When the protein is expressed in a recombinant system, seleno-methionine (Se-Met) protein crystals can be used for heavy atom phasing. Replacing the Met residues of a given protein with Se-Met results in good anomalous diffraction signals from the selenium. Recently, halide soaks have been successful in derivatizing protein crystals [66].

The refined heavy atom positions and parameters are used as the starting set in the calculation of the experimental phases. Once experimental phases are obtained, an electron density map can be calculated by Fourier transform procedures. The electron density map is the most valuable source of information, and eventually leads to model building. If the starting experimental phases are of poor quality and the maps are uninterpretable, the phases can be improved by density modification procedures, using programs such as DM or SOLOMON [67], which also include solvent flattening and noncrystallographic symmetry averaging.

Model building and refinement

Model building and refinement is the process of constructing a 3D molecular structure to fit the experimental electron density and, at the same time, maintaining reasonable stereochemistry. This is achieved using either interactive computer graphics programs, such as Xfit and O [68,69], or automated fitting programs, such as ARP/wARP and QUANTA [70,71]. With minimal or no user intervention, ARP/wARP can automatically build and refine a protein model, starting with diffraction data to a resolution of 2.3 Å or higher, whereas QUANTA combines graphics and automated model building into a single program suite. On completion of model building, the model is refined, using programs that include REFMAC, XPLOR (CNX or CNS) and TNT [63,67,72]. Automated model validation can be done using WHAT-CHECK and SFCHECK [67,73].

In protein crystallography for pharmaceutical development, often of greater interest than the structure of the target protein is the structure of the protein in complex with an inhibitor or drug lead. A large number of protein–ligand complex structures might be required in the iterative process of structure-based drug design. The development of HIV-protease inhibitors that are currently used in the treatment of AIDS was greatly facilitated by the protein crystallographic determination of HIV-protease–ligand complex structures. There are currently >40 HIV-protease complex structures deposited in the PDB, and it is probable that this number does not represent all of the HIV-protease complex structures that were determined in the development of these AIDS drugs.

Because protein–ligand complexes often crystallize in the same crystal form as the apoprotein, the computational challenge in determining protein–ligand complex structures is not so much in determining the structure of tens of thousands of kDa of protein, but rather a few hundred Da of ligand. In high-throughput structure determination of protein–ligand complexes (‘crunching complexes’), tools that can locate, build and refine the structure of the bound ligand with minimal human intervention are desired. One of the best tools for automated ligand-fitting is X-LIGAND, part of the QUANTA package from Accelrys (San Diego, CA, USA). X-LIGAND automatically searches unoccupied regions of electron density and tries to fit in the structure of a ligand [74]. The conformational space of the ligand can be searched for each region of fitted density, and then the best fit conformation of the ligand can be refined by real-space torsional angle refinement.

Conclusions

With the expected exponential growth in the number of protein targets that were identified as a result of the completed Human Genome Project, high-throughput systems are necessary for protein expression, protein purification, crystallization and structure determination. High-speed structure determination of protein–ligand complexes will be demanded by the need to know exactly how a ligand binds to its target. The development of integrated technology platforms and database systems to automate all steps – from clone to final structure – is extremely important for realizing the full impact that high-throughput crystallography can have on the drug design process. The combination of crystallographically validated protein–ligand structures with increasingly sophisticated computational chemistry tools offers an accelerated model to pursue a growing number of protein targets for small-molecule drug discovery.

Acknowledgements

We wish to thank John Wooley of the University of California at San Diego (CA, USA), John Mulligan of Blue Heron BioTechnology (Bothell, WA, USA), and G. Sridhar Prasad at Syrrx (San Diego, CA, USA) for providing information for this review. We also acknowledge contributions to this work by Hans-Thomas Richter and thank the research and support staff of Emerald BioStructures.

References

1 Mundy, C. (2001) The Human Genome Project: a historical perspective. *Pharmacogenomics* 2, 37–49

2 Debouck, C. and Metcalf, B. (2000) The impact of genomics on drug discovery. *Annu. Rev. Pharmacol. Toxicol.* 40, 193–207

3 Mittl, P.R. and Grutter, M.G. (2001) Structural genomics: opportunities and challenges. *Curr. Opin. Chem. Biol.* 5, 402–408

4 Hol, W.G. (2000) Structural genomics for science and society. *Nat. Struct. Biol.* 7, 964–966

5 Service, R.F. (2000) Structural genomics offers high-speed look at proteins. *Science* 287, 1954–1956

6 Williamson, A.R. (2000) Creating a structural genomics consortium. *Nat. Struct. Biol.* 7 (Suppl.), 953

7 Yokoyama, S. *et al.* (2000) Structural genomics projects in Japan. *Nat. Struct. Biol.* 7, 943–945

8 Lewis, H.A. *et al.* (2001) A structural genomics approach to the study of quorum sensing: crystal structures of three LuxS orthologs. *Structure* 9, 527–537

9 Vitkup, D. *et al.* (2001) Completeness in structural genomics. *Nat. Struct. Biol.* 8, 559–566

10 Krogh, A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580

11 Muller, G. (2000) Towards 3D structures of G protein-coupled receptors: a multidisciplinary approach. *Curr. Med. Chem.* 7, 861–888

12 Yoshida, M. *et al.* (2001) Proteomics as a tool in the pharmaceutical drug design process. *Curr. Pharm. Des.* 7, 291–310

13 Gane, P.J. and Dean, P.M. (2000) Recent advances in structure-based rational drug design. *Curr. Opin. Struct. Biol.* 10, 401–404

14 Langer, T. and Hoffmann, R.D. (2001) Virtual screening: an effective tool for lead structure discovery? *Curr. Pharm. Des.* 7, 509–527

15 Verkhivker, G.M. *et al.* (2000) Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput. Aided. Mol. Des.* 14, 731–751

16 Zeng, J. (2000) Mini-review: computational structure-based design of inhibitors that target protein surfaces. *Comb. Chem. High Throughput Screen.* 3, 355–362

17 Hassell, A.M. *et al.* (1994) Two distinct approaches to crystallization results-recording databases. *Acta Cryst.* D50, 459–465

18 Gilliland, G.L. *et al.* (1996) The Biological Macromolecule Crystallization Database and NASA Protein Crystal Growth Archive. *J. Res. Natl. Inst. Stand. Technol.* 101, 309–320

19 Gopalakrishnan, V. *et al.* (2000) Intelligent aids for parallel experiment planning and macromolecular crystallization. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8, 171–182

20 Stewart, L. and Kim, H. (2000) Crystal Monitor: The relational database application for crystal growth. *American Crystallographic Association Annual Meeting*, 22–27 July 2000, St. Paul, MN, USA (Abstract W0140)

21 Jackob, C. *et al.* (2001) Automating the Crystallography Laboratory for Structure-based Drug Design. *American Crystallographic Association, Annual Meeting*, 21–26 July 2001, Los Angeles, CA, USA (Abstract W0370)

22 Hennessy, D. *et al.* (2000) Statistical methods for the objective design of screening procedures for macromolecular crystallization. *Acta Cryst.* D56, 817–827

23 Jurisica, I. *et al.* (2001) Intelligent decision support for protein crystal growth. *IBM Systems Journal* 40, 394–409

24 Segelke, B.W. (2001) Efficiency analysis of screening protocols used in protein crystallization. *J. Cryst. Growth* 232, 553–562

25 Bertone, P. *et al.* (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.* 29, 2884–2898

26 Strausberg, R. *et al.* (1999) The Mammalian Gene Collection. *Science* 286, 455–457

27 Boguski, M. *et al.* (1993) dbEST-database for ‘expressed sequence tags’. *Nat. Genet.* 4, 332–333

28 Salghetti, S.E. *et al.* (2000) Functional overlap of sequences that activate transcription and signal ubiquitin-mediated proteolysis. *Proc. Natl. Acad. Sci. U. S. A.* 97, 3118–3123

29 Edwards, A.M. *et al.* (2000) Protein production: feeding the crystallographers and NMR spectroscopists. *Nat. Struct. Biol.* 7 (Suppl.), 970–972

30 Davies, A.H. (1994) Current methods for manipulating baculoviruses. *Biotechnology (NY)* 12, 47–50

31 Albala, J.S. *et al.* (2000) From genes to proteins: high throughput expression and purification of the human proteome. *J. Cell. Biochem.* 80, 187–191

32 Cregg, J.M. *et al.* (2000) Recombinant protein expression in *Pichia pastoris*. *Mol. Biotechnol.* 16, 23–52

33 Kigawa, T. *et al.* (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. *FEBS Lett.* 442, 15–19

34 Cho, H.S. *et al.* (2001) *In vitro* protein production for structure determination with the rapid translation system (RTS). *Roche Applied Science* Application Note No. 4/2001, Rapid Translation System RTS, 500, pp. 501–504

35 Porfiri, E. *et al.* (1995) Purification of Baculovirus-Expressed Recombinant Ras and Rap Proteins. *Methods Enzymol.* 225, 13–21

36 Lucast, L.J. *et al.* (2001) Large-scale purification of a stable form of recombinant tobacco etch virus protease. *Biotechniques* 30, 544–546

37 Hong, L. *et al.* (2000) Structure of the protease domain of memapsin 2 (β-secretase) complexed with inhibitor. *Science* 290, 150 –153

38 Bernstein, B.E. *et al.* (1998) The importance of dynamic light scattering in obtaining multiple crystal forms of *Trypanosoma brucei* PGK. *Protein Sci.* 7, 504–507

39 McPherson, A. (1992) Two approaches to the rapid screening of crystallization conditions. *J. Cryst. Growth* 122, 161–167

40 Chayen, N.E. (1998) Comparative studies of protein crystallization by vapour-diffusion and microbatch techniques. *Acta Cryst.* D54, 8–15

41 Kühlbrandt, W. (1992) Two-dimensional crystallization of membrane proteins. *Quart. Rev. Biophys.* 25, 1–49

42 Michel, H. (1983) Crystallization of membrane proteins. *Trends Biochem. Sci.* 8, 56–59

43 Nollert, P. *et al.* (2002) Crystallization of membrane proteins *in cubo*. *Methods Enzymol.* 343, 183–199

44 Landau, E.M. and Rosenbusch, J.P. (1996) Lipidic cubic phases: a novel concept for the crystallization of membrane proteins. *Proc. Natl. Acad. Sci. U. S. A.* 93, 14532–14535

45 Luecke, H. *et al.* (1999) Structure of bacteriorhodopsin at 1.55 Å resolution. *J. Mol. Biol.* 291, 899–911

46 Kolbe, M. *et al.* (2000) Structure of the light-driven chloride pump halorhodopsin at 1.8 Å resolution. *Science* 288, 1390–1396

47 Luecke, H. *et al.* (2001) Crystal structure of sensory rhodopsin II at 2.4 angstroms: insights into color tuning and transducer interaction. *Science* 293, 1499–1503

48 Ermler, U. *et al.* (1994) Structure of the photosynthetic reaction centre from Rhodobacter sphaeroides at 2.65 Å resolution: cofactors and protein-cofactor interactions. *Structure* 2, 925–936

49 Lancaster, C.R. *et al.* (2000) Structural basis of the drastically increased initial electron transfer rate in the reaction center from a *Rhodopseudomonas viridis* mutant described at 2.00 Å resolution. *J. Biol. Chem.* 275, 39364–39368

50 McLuskey, K. *et al.* (2001) The crystallographic structure of the b800-820 lh3 light-harvesting complex from the purple bacteria *Rhodopseudomonas acidophila* strain 7050. *Biochemistry* 40, 8783–8789

51 Goodwill, K.E. *et al.* (2001) High-throughput X-ray crystallography for structure-based drug design. *Drug Discov. Today* 6, S311–S118

52 Stevens, R.C. (2001) New paradigms in drug discovery chemistry. In *Drug Discovery By Design 2001*, IBC

53 Henry, C.M. (2001) Structure-based drug design. *Chem. Eng. News* 79, 69–74

54 Stewart, L. (2001) High-throughput crystallization and X-ray structure determination of protein–ligand complexes. In *Drug Discovery by Design 2001*, IBC

55 Kim, H. and Stewart, L. (2000) Crystallization Tray (US Patent 6,039,804), Emerald BioStructures, Bainbridge Island, WA, USA

56 Nienaber, V.L. *et al.* (2000) Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nat. Biotechnol.* 18, 1105–1108

57 Bernstein, B.E. *et al.* (1997) Synergistic effects of substrate-induced conformational changes in phosphoglycerate kinase activation. *Nature* 385, 204–205

58 Bernstein, B.E. and Hol, W.G. (1998) Crystal structures of substrates and products bound to the phosphoglycerate kinase active site reveal the catalytic mechanism. *Biochemistry* 37, 4429–4436

59 Stewart, L. *et al.* (2001) Topotecan bound to human topoisomerase I at 2.0 angstrom resolution. In *92nd Annual Meeting of the American Association for Cancer Research*, 24–28 March 2001, New Orleans, LA, USA (Vol. Proceedings Supplement), pp. 80

60 Muchmore, S.W. *et al.* (2000) Automated crystal mounting and data collection for protein crystallography. *Structure Fold Des.* 8, R243–R246

61 Kissinger, C.R. *et al.* (1999) Rapid automated molecular replacement by evolutionary search. *Acta Cryst.* D55, 484–491

62 Navaza, J. (1994) AMoRe: an automated package for molecular replacement. *Acta Cryst.* A50, 157–163

63 Brunger, A.T. *et al.* (1998) Crystallography and NMR system: a new software system for macromolecular structure determination. *Acta Cryst.* D54, 905–921

64 Kissinger, C.R. and Smith, B.A. (2001) Automated Molecular Replacement. *American Crystallographic Association Annual Meeting,* 21–26 July 2001, Los Angeles, CA, USA (Abstract WO299)

65 Mixon, M. *et al.* (2001) High-throughput molecular replacement on a Linux cluster. In *Western Canadian Structural Biology Workshop and Meeting: High-throughput crystal structure determination for structural genomics,* 18–20 October 2001, Banff, Alberta, Canada

66 Dauter, Z. *et al.* (2000) Novel approach to phasing proteins: derivatization by short cryo-soaking with halides. *Acta Cryst.* D56 (Pt 2), 232–237

67 Collaborative Computational Project, N. (1994) The CCP4 Suite: programs for protein crystallography. *Acta Cryst.* D50, 760–763

68 McRee, D.E. (1993) *Practical Protein Crystallography*. p. 290, Academic Press

69 Jones, T.A. *et al.* (1991) Improved methods for binding protein models in electron density maps and the location of errors in these models. *Acta Cryst.* A47 (Pt 2), 110–119

70 Lamzin, V.S. and Wilson, K.S. (1993) Automated refinement of protein models. *Acta Cryst.* D49, 129–147

71 Oldfield, T.J. (2001) A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent. *Acta Cryst.* D57 (Pt 1), 82–94

72 Tronrud, D.E. (1997) TNT refinement package. *Methods Enzymol.* 277, 306–319

73 Vriend, G. (1990) WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* 8, 52–56

74 Oldfield, T.J. (2001) X-LIGAND: An application for the automated addition of flexible ligands into electron density. *Acta Cryst.* D57, 696–705