TUTORIAL REVIEW

CSR
www.rsc.org/csr

High-throughput protein crystallography and drug discovery

Ian Tickle,* Andrew Sharff, Mladen Vinković, Jeff Yon and Harren Jhoti
Astex Technology Ltd., 436 Science Park, Milton Road, Cambridge, UK CB4 0QA

Received 30th January 2004
First published as an Advance Article on the web 20th September 2004

Single crystal X-ray diffraction is the technique of choice for studying the interactions of small organic molecules with proteins by determining their three-dimensional structures; however the requirement for highly purified protein and lack of process automation have traditionally limited its use in this field. Despite these shortcomings, the use of crystal structures of therapeutically relevant drug targets in pharmaceutical research has increased significantly over the last decade. The application of structure-based drug design has resulted in several marketed drugs and is now an established discipline in most pharmaceutical companies. Furthermore, the recently published full genome sequences of *Homo sapiens* and a number of micro-organisms have provided a plethora of new potential drug targets that could be utilised in structure-based drug design programs. In order to take maximum advantage of this explosion of information, techniques have been developed to automate and speed up the various procedures required to obtain protein crystals of suitable quality, to collect and process the raw X-ray diffraction data into usable structural information, and to use three-dimensional protein structure as a basis for drug discovery and lead optimisation.

This *tutorial review* covers the various technologies involved in the process pipeline for high-throughput protein crystallography as it is currently being applied to drug discovery. It is aimed at synthetic and computational chemists, as well as structural biologists, in both academia and industry, who are interested in structure-based drug design.

Ian Tickle

Dr Ian Tickle is the Director of X-ray Technology at Astex Technology. Prior to joining Astex in 2001, he was a Senior Research Fellow in the School of Crystallography, Birkbeck College, University of London. His research interests there included the development of methods and software for structure solution and refinement of macromolecular structures by X-ray crystallography and electron microscopy, and the analysis of correlated atomic motions in protein structures. Dr Tickle graduated with a degree in Chemistry from the University of Oxford in 1969 and also received his D.Phil. in Crystallography in 1972 from Oxford University.

Andrew Sharff

Dr Andrew Sharff is a Senior Research Associate at Astex Technology. Before joining Astex in 2000, he worked with Professor Florante Quiocho at the Howard Hughes Medical Institute in Houston, USA, Professor Steven Neidle at the Institute of Cancer Research in London and with Dr Ben Luisi at Cambridge University. Dr Sharff graduated with a degree in Biochemistry from the University of Bristol in 1987 and received his D.Phil. in Molecular Biophysics in 1991 from Oxford University.

Mladen Vinković

Dr Mladen Vinković is Senior Research Associate at Astex Technology. Before joining Astex he was head of Molecular Design at PLIVA, Croatia, and held the position of Assistant Professor of Bioinorganic Chemistry at the University of Zagreb. He has also worked in Prof. Sir Tom Blundell's groups in London and Cambridge. Dr Vinković received his PhD in Chemistry from the University of Zagreb in 1994. He has co-authored 40 patents and scientific publications on powder, small molecule and protein diffraction, molecular modelling, design of NCE and drug polymorphism.

Jeff Yon

Dr Jeff Yon is the Director of Protein Technology at Astex Technology. Before joining Astex in 2001 he worked at the Janssen Research Foundation in Beerse, Belgium (a Johnson and Johnson company) and at Pfizer's laboratories in Sandwich, producing recombinant proteins for structural biology and screening. Dr Yon graduated with a degree in Biochemistry from the University of Cambridge in 1982 and a PhD in Genetics from Birmingham University in 1986.

DOI: 10.1039/b314510g

 This journal is © The Royal Society of Chemistry 2004

Introduction

The concept of high-throughput crystallography was first proposed in the mid-1990's. It was inspired by the significant achievements in the genomics arena in which the DNA of several organisms, including *Homo sapiens*, were being sequenced using high-throughput methodologies.[1] Over the last 15 years the broader application of 'high-throughput' methodologies has become commonplace within the pharmaceutical industry with the advent of high-throughput chemistry and screening.[2] It was during this period that protein crystallography was initially unable to keep pace and so the other drug discovery technologies being performed in a high-throughput mode became the focus of many pharmaceutical companies.

More recently there has been a resurgence of interest in using structure-based approaches, driven largely by major technology developments in crystallography, which has resulted in many new crystal structures of therapeutics targets. Many of these technology advances have been pioneered by the 'structural genomics' initiatives that were set up with the aim of solving crystal structures of representatives from protein families for which little or no structural information was known.[3] The original goal of the structural genomics projects was to obtain crystal structures for all known protein families and most have focused on the various bacterial genomes.

The ability to rapidly obtain crystal structures of a target protein in complex with small molecules has further increased the impact of protein crystallography in drug discovery. Indeed, in some pharmaceutical companies up to a third of all lead optimisation programs utilise information from protein crystal structures. The increase in the rate of obtaining crystal structures of protein–ligand complexes has now allowed X-ray crystallography to be used for lead discovery and, in particular, as the method of choice for fragment-based screening approaches.

Protein production

The quality criteria for proteins for crystallisation are stringent: it is generally accepted that samples should be pure ($>$ 95% pure) and both conformationally and chemically homogeneous. One of the key challenges for high-throughput approaches is to generate multiple samples meeting these high standards. Producing such samples is a bottleneck in conventional structural biology laboratories, and can become limiting for large-scale structural genomics efforts.[4] Recombinant protein production involves many steps: cloning of a suitable DNA sequence, generation of expression constructs, testing and optimisation of gene expression, scale-up, purification and characterisation. Over the last 2–3 years preferred methods for these steps have begun to emerge, as reviewed below. In general these methods rely on using low sample volumes for as many of the cloning and expression steps as possible, both for ease of handling multiple samples and to reduce costs, and on the use of robust, automatable, parallel processes. Protein purification also makes heavy use of affinity tags fused to the protein of interest, to allow generic purification schemes.[5]

One of the first questions to be addressed for a given target is the choice of expression system. The early phases of the public structural genomics programs have focused on developing high-throughput methods using *E. coli* as the expression host (see below). While the system has many advantages in terms of cost, speed and ease of use, it has become clear that the success rate for soluble *E. coli* expression of eukaryotic proteins is low.[4,6] As a result, several eukaryotic expression systems have also been explored for high-throughput approaches. There are reports of expression studies using the yeasts *S. cerevisiae* and *P. pastoris* in microplate format, which allows many samples to be processed in parallel. Perhaps most interesting from a drug discovery perspective is the development of microplate-based growth and expression-screening methods for the baculovirus/insect cell system.[5] This system has a good record of generating mammalian proteins for structural biology applications, and has been a system of choice for production of protein kinases, a prominent drug target class over the last 5–10 years.

As discussed above, most high-throughput approaches are currently based on expression in *E. coli.* The first step in the protein production process, namely obtaining a suitable DNA sequence, is often straightforward, as the available cDNA collections have become more comprehensive in recent years. If no characterised cDNA is available, PCR cloning from cDNA has proven to be a robust cloning method. The next step is subcloning of the desired regions of the cDNA into expression vectors. At this stage judgements are made about what is likely to constitute the best protein for crystallisation. As various construct options will usually be explored (*e.g.* where the N- and C-terminal boundaries should be, which purification tag should be used *etc.*), a range of expression constructs is generally made and tested in parallel. Recombination-based cloning systems such as Gateway™ (Invitrogen, Carlsbad, CA, USA) or Creator™ (Clontech, Palo Alto, CA, USA) are increasing in popularity for these sub-cloning steps due to their simplicity and fidelity. Furthermore, as the reactions consist of a series of liquid addition steps, recombination-based subcloning can readily be automated.

When screening *E. coli* constructs for protein expression, several parameters are usually varied. Different strains will be employed, as this can have a dramatic effect on expression. Induction of expression may be tested at different temperatures, as lowering the expression temperature may improve the yield of soluble protein. Different growth medium compositions are often explored; this is of particular importance in high-throughput structure determination approaches, as a medium suitable for selenomethionine labelling for SAD/MAD phasing may be required (see 'Structure determination by *ab initio* phasing methods' below). Since each construct will be tested in several different conditions, and a number of constructs will have been made, there may be very many expression trials to conduct. It is therefore no surprise that expression screening is often carried out using small sample volumes in microplates, with only the most promising constructs or

Dr Harren Jhoti is the Chief Scientific Officer and a Founder of Astex Technology, a UK-based biotechnology company with around 100 staff. Before setting up Astex in 1999, he led the Structural Biology and Bioinformatics groups at GlaxoWellcome (1991–1999), applying protein structure analysis to drug discovery. While at GlaxoWellcome, he was involved in structure-based drug design projects aimed at a variety of therapeutic targets including blood coagulation proteases, viral proteases, kinases and other signal transduction proteins. Dr Jhoti graduated with a degree in Biochemistry from the University of London in 1985 and received his PhD in Protein Crystallography in 1989, from Birkbeck College, University of London.

Harren Jhoti

conditions taken forward for scale-up. The readout of expression screening may be the production of a visible band on SDS-PAGE, or production of the protein may be detected *via* antibody binding to the protein or affinity tag.[5,7]

Once the most promising constructs have been identified, expression is scaled up and the protein is purified to homogeneity. Protein purification for high-throughput approaches is often based on affinity tags, which allow standard purification schemes to be used. These methods can be run with a minimum of operator involvement for several samples sequentially using commercially available automated chromatography systems (such as those from Amersham Biosciences, http://www.amershambiosciences.com, or Applied Biosystems, http://home.appliedbiosystems.com). Alternatively home-built robotic purification systems (such as those at the Genetics Institute of the Novartis Foundation, http://web.gnf.org) may be used. In some cases, however, additional purification steps will be required to achieve the level of purity suitable for crystallisation studies;[7] if these must be developed for each target they will reduce throughput.

Whilst the majority of high-throughput structural biology programs have used *E. coli* or one of the systems described above, several alternative approaches are also being tested. Cell-free expression (in which a nucleic acid template is added to a protein-synthesising lysate) is receiving significant attention, as recent developments have improved yields. This system may offer a route to production of difficult targets, for example proteins that are toxic in bacteria. The use of reporter gene fusions has also resulted in several published structures. In this method the protein of interest is fused to a reporter, often green fluorescent protein, so that soluble expression of the target gives rise to an increased signal from the reporter protein. If a library of target protein variants is constructed, the reporter can be used to select for variants with improved solubility, which then permit purification and crystallisation. It is not yet clear, however, what impact either cell-free expression or reporter gene methods will have on productivity in structural biology.[8]

In summary, methods for protein production for high-throughput structure determination have largely focused on *E. coli* expression. Low-volume, high-density microplate formats have been used for the cloning and expression screening stages of the process. Automation has been applied, either in the form of commercially available liquid handling systems or purpose-built robots. This has resulted in greater throughput both in cloning and in protein expression. Obtaining soluble protein remains a bottleneck that is being tackled through the use of different expression hosts or alternative expression approaches. Examination of the outputs of structural genomics programs (Target DB: http://targetdb.pdb.org) indicates that over five times as many target proteins were listed as 'purified' during 2003 compared with 2002, providing strong evidence of improvements in protein production.

Crystallisation

The crystallisation process has traditionally been considered as a major bottleneck in protein X-ray crystallography (PX). It can be divided into two logical steps: screening for initial crystallisation conditions and optimisation of conditions to produce diffraction quality single crystals. The availability of limited amounts of sample, combined with the huge space of crystallisation parameters, makes the discovery of initial crystallisation conditions challenging. However, this field has been revolutionised recently by developments in automation, miniaturisation and process integration. Current automation in protein crystallisation takes advantage of commercially available liquid handling systems initially developed for other applications such as high-throughput screening (HTS), as well as developments of high density crystallisation micro-plates containing 96 or more reservoirs and compliant with the SBS (Society for Biomolecular Screening) standard.

Several different parameters can be altered to entice protein molecules to nucleate and form a protein crystal, such as ionic strength, precipitant concentration, additives, pH and temperature. As a consequence, a number of different crystallisation techniques have been developed,[9] but only a few of them are suited to high-throughput automation. Vapour diffusion methods are still the most popular, with the sitting drop technique currently more popular than the more traditional hanging drop approach, as it is simpler to automate. The micro-batch method of crystallisation under oil is more straightforward to automate and its use is increasing.[10]

A typical high-throughput crystallisation screening campaign starts with automated preparation and/or reformatting of screening (precipitant) solutions into crystallisation micro-plates. Thousands of crystallisation experiments can then be set up by mixing screening and protein solutions using robotic systems in which a wide range of variables are explored. The resulting large number of crystallisation drops can be monitored regularly by an imaging robot and collected images can then either be analysed manually on a computer screen or by image recognition software.

Some researchers believe that the crystallisability of a protein sample can be assessed by using only 50–100 crystallisation conditions. However, there are examples of protein samples that produce crystals in only one of thousands of screening conditions[10] and moreover the same screening conditions may not produce crystals every time. Therefore, most groups rely on more extensive screening, typically using from 300–3000 conditions. In order to minimise resource usage, many laboratories will start with 300–400 conditions and then only if these are unsuccessful, gradually introduce additional ones. More than 1500 different crystallisation screening solutions are available commercially from various manufacturers (see for example Hampton Research, http://www.hamptonresearch.com), and customised screens can also be designed in-house. The 'Matrix Maker' robot from Emerald Biostructures (http://www.decode.com/emeraldbiostructures) is specifically designed for high-throughput crystallisation solution preparation. It can simultaneously handle 48 or more stock solutions in continuous dispensing mode. Eight-channel liquid handling systems such as 'Genesis' from Tecan or 'Microlab Star' from Hamilton (http://www.hamilton.ch) are used either to reformat solutions into crystallisation plates[11] or to prepare them from stock solutions on a smaller scale. To achieve higher throughput some laboratories are using 96-channel robots such as the Robbins 'Hydra' for reformatting.[12]

If the amount of protein is limited, a new approach, referred to as 'nanocrystallogenesis', can be employed in which very small crystallisation drops are used, containing as little as 25–100 nl of protein.[13] The wave of interest in nanocrystallogenesis has been initiated by the availability of solenoid valve-based nanodispensers. The current Cartesian line of products from Genomic Solutions (http://www.genomicsolutions.com) is capable of dispensing drops from 20 nl to several μl in sitting or hanging drop plates ('Honeybee' systems), as well as dispensing drops through the oil in the micro-batch method. Critically, these machines are able to dispense very viscous crystallisation solutions such as 30% PEG8K with high accuracy. The other proven technology for nano-dispensing protein crystallisation drops uses positive displacement nano-tips and is implemented in the TTP LabTech 'Mosquito' robot (http://www.ttplabtech.com).

Several manufacturers of micro-plates have developed special automation-friendly protein crystallisation micro-plates. Most

of them are designed for the sitting-drop vapour diffusion method. For example, Greiner's 'CrystalQuick' plate features 96 reservoirs for the screening solution with either one or three crystallisation wells per reservoir. The Coming 'CCP384' plate has 192 reservoirs for screening solutions with one clear flat-bottomed crystallisation well per reservoir. For the micro-batch under oil method, Greiner developed the '1536IMPACT' micro-plate containing 1536 conical wells. Because the micro-batch method does not require a reservoir for the screening solution, a much high density of crystallisation experiments on one microplate as compared with the vapour diffusion method is possible.

Visualisation robots such as 'Minstrel III' (Robodesign International, http://www.robodesign.com) or 'Rhombix Vision' (DataCentric Automation, http://www.dcacorp.com) for scanning crystallisation micro-plates are in many cases linked with plate hotels/incubators to enable high-throughput unattended periodical image acquisition. The 'Rhombix Vision' robot can collect images using various illumination schemes such as bright field, dark field and polarised light with several angles of polarisation, in order to maximise the contrast between crystal and background. Although there have been significant advances in the development of crystal recognition algorithms and software, at present many users trust it only to eliminate clear and other drops that the software can reliably classify as negatives. The remaining images are then inspected manually.

There are several examples where the whole protein crystallisation process, from screening solution preparation to image classification, has been integrated into one large hands-free automated system. However, mainly due to cost effectiveness, most laboratories tend to automate only parts of the process and leave it to operators to bridge the gaps, such as plate transfer between various robots.

A recent development in protein crystallisation has been the introduction of disposable micro-fluidic crystallisation chips with Fluidigm's 'Topaz' system (http://www.fluidigm.com). This uses the free-interface diffusion method at low nano-litre scale (25 nl). Although this equipment cannot match the throughput of automated vapour diffusion and micro-batch methods, the free-interface diffusion method explores more of the crystallisation space in each experiment than vapour diffusion and has claimed several successes.[14]

Once the first crystallisation conditions have been found, there is still a lot of work involved in optimising the crystallisation conditions to produce diffraction quality crystals. In many cases the same automation equipment as used for the initial screening is applied in the optimisation step. Some laboratories even continue using nano-drops in crystallisation optimisation. However, a significant fraction of the crystal forms cannot be optimised to diffract to a useful resolution, and a redesign of the protein sequence is often necessary. This problem is statistically highlighted by the various structural genomics initiatives, and changes the scope from a search for crystals to a search for well-diffracting crystals.

X-Ray data collection

The steady advance of technology, largely geared to synchrotron beamlines, has dramatically increased the speed and ease of X-ray data collection over the past decade.[15] Until recently, the applications of these technological advances have been carried out in relative isolation. For instance, new third-generation synchrotrons have provided more intense X-rays, new designs of X-ray optics have given brighter, cleaner, more controllable and better collimated X-ray beams, and new faster, larger and more sensitive X-ray detectors have allowed higher quality data to be collected much more rapidly.

However, the demands for even greater productivity from structural genomics and the ever increasing application of protein crystallography to drug discovery projects have placed further burdens on X-ray data collection throughput. Indeed, the success of fragment-based screening by X-ray crystallography (see final section) is critically dependent on the ability to collect good data rapidly. The requirement for high-throughput structure solution has highlighted the manual nature of X-ray data collection and processing methodologies and identified these as significant bottlenecks in the structure determination pipeline.

The solution to overcoming these barriers lies in the automation of both data collection and processing.[16] The rate-limiting step in high-throughput data collection at synchrotrons is often the manual intervention required to mount and align crystals. On third generation sources, this can easily exceed half the time taken to collect the data, a serious loss of data-collection efficiency. Screening to find the best crystal for collection from within a set is also very time consuming and inefficient and is a task ideally suited for automation. The development of automated sample changers was a direct response to the challenge to eliminate these sources of inefficiency, although the first such system was not developed at a synchrotron, but by Abbott Laboratories on their in-house X-ray system.[17] Many synchrotrons, both in the USA and in Europe, have now developed and installed their own systems, and new synchrotron beamlines have automation included as a specific requirement of their design. The influence of sample changers in increasing data collection throughput and efficiency is now widely recognised, to the point that several automated sample changers are now commercially available, not only for installation on synchrotron beamlines but also for in-house X-ray generators.

Developments in technology have not been limited solely to synchrotron sources. The latest generation of high intensity X-ray generators, coupled with improved X-ray optics have revolutionised the laboratory X-ray system, to the point where complete, integrated in-house systems, such as those from Rigaku MSC (http://www.rigakumsc.com) and Bruker AXS (http://www.bruker-axs.com), are available which can now rival some synchrotron beamlines for X-ray intensity. Coupled with automated sample changers, high-throughput X-ray crystallography is now possible in the laboratory. Furthermore, recent developments in X-ray optics from Rigaku MSC have made in-house *ab initio* structure solution by SAD (single anomalous diffraction) sulfur phasing possible, thus obviating the need, in suitable cases, for time-consuming heavy-atom methods.[18]

Advances in and automation of data collection hardware have to be matched by parallel development in the software required to control data collection and processing. Until recently, software to control both synchrotron beamlines and in-house X-ray detectors has been written in virtually complete isolation from software for data analysis and processing. The result is a sub-optimal system for experimental design and control. Initial set-up of data collection has to be carried out using the beamline/detector control software, however crystal characterisation is determined separately, using the data processing software (often on a separate computer at synchrotron beamlines). A decision is made on whether or not the diffraction quality of the crystal is sufficient for the desired outcome and if so, the design of the data collection experiment then has to be determined manually from the results of the crystal characterisation. This information has then to be fed back to the control software to initiate data collection – until recently the control systems installed at the beamlines have not allowed for any direct interaction between the control and processing software. Finally, integration, scaling and reduction

of the final data are performed offline using the processing software.

This manual interplay between initial screening, crystal characterisation, data collection and post-processing is slow and inefficient. With data-collection times falling, the time taken to set up data collection can take up a significant proportion of the total time for the whole experiment. The goal for software automation is an integrated, 'smart' system that can encompass control of data collection, experimental design/set-up and post-processing with a minimum of human input. Such an 'expert' system is vital for a fully efficient implementation of automatic sample changers on beamlines. The first generation of such a system, called Blu-Ice,[19] incorporates a graphical user interface for beamline set-up and control, and has already been implemented at several synchrotron sources in the US. Leslie *et al.*[20] describe a much more sophisticated expert system, fully integrating data processing and detector control. This system will allow automatic data analysis and intelligent decision making as to whether the crystal is suitable for data collection, and on this basis will then determine the optimum experimental design, and then collect and process the data. Such software is also being developed on in-house X-ray systems. These not only control the data collection itself, but also process the data 'on the fly' as it is collected, thus providing the researcher with fully-processed data in the minimum possible time.

Structure determination by *ab initio* phasing methods

The central problem in macromolecular structure determination by X-ray crystallography is solution of the phase problem. X-ray data collected from a protein crystal consist of structure factor amplitudes, however a critical component, namely the phase associated with each amplitude, cannot be recorded directly. In order to determine the structure, this phase information has to be recovered. There are three principal methods for determining the phases. Single/multiple isomorphous replacement (SIR/MIR) and single/multiple anomalous diffraction (SAD/MAD) are *ab initio* methods, for which no prior structural knowledge is required. The molecular replacement (MR) method on the other hand relies on having available a structurally similar model of the target protein.

Traditionally, structure solution, especially using *ab initio* methods, and model building have been largely manual, step driven and frequently non-linear processes. Until recently this had not been a major issue, as the throughput of data for new structures had been fairly low and therefore structure solution, though both manual and laborious, did not represent a significant bottleneck.

However, with the demands of a high-throughput structure solution process, driven by the success of the various structural genomics projects in crystallising many new and novel proteins, together with the advent of fragment-based screening by X-ray crystallography as a viable technology for drug discovery, this is no longer the case. The need for speed provides an imperative to automate structure determination, from phase determination, through to model building and refinement.[21–24] It should be noted that although *ab initio* structure solution methods themselves do not play an important role in fragment-based screening for drug discovery, as the majority activity involves solving structures of protein–ligand complexes based on the known apo structure, there is nevertheless a demand for automated model building/modification techniques.

Ab initio methods for solving the phase problem are very well established, albeit with a tendency to require specialised expertise, and hence do not lend themselves easily to automation.[25] The most popular methods for structure determination, MIR and MAD, both require the ordered introduction of appropriate heavy or anomalous scatterers into the protein crystal. Although the techniques for achieving these ends are well understood, they have several drawbacks, particularly non-isomorphism (changes in the crystal structure of the protein induced by the heavy atoms) and the time taken for derivative preparation. New approaches to phasing, such as fast halide soaks and SAD on sulfurs and other endogenous weak anomalous scatterers, although not 'automatic' are much simpler and faster than traditional techniques and are growing in popularity.[25]

Software programs for location of heavy atom sites/anomalous scatterers by Patterson interpretation or application of Direct Methods techniques, as well as phase improvement techniques, such as solvent flattening/flipping have also developed considerably in the past decade. Programs such as SHELX, SOLVE, SHAKE & BAKE and SHARP have greatly simplified the task of finding heavy atom sites and generating accurate phases from them, and have speeded up the whole process of structure determination.[26] However, on their own, they do not go far enough in providing for total automation. Extensions of some of these packages, for instance SOLVE/RESOLVE, AUTOSHARP, BnP and CHART have gone a long way to incorporate automation, by further simplifying and streamlining the whole process. These packages have a simple set-up requiring minimum input and completely automate the location of heavy atom/anomalous scattering sites, phase determination and phase improvement by solvent flattening/flipping.

Structure determination by molecular replacement

When an approximate structural model of a protein under investigation is available, either from NMR, from an X-ray structure of a homologous protein or from homology modelling, initial phases can be obtained by the molecular replacement (MR) method using the approximate structure as a search model. With an increasing number of new protein structures solved and deposited with the PDB (Protein Data Bank, http://www.rcsb.org/pdb), there is an ever-increasing chance that one of the domains of the target protein will have a previously structurally-characterised homologue. Where there are only one or two protein molecules per asymmetric unit of the crystal, and where the search molecule is structurally similar to the target protein, the MR method is fairly straightforward and fast with programs such as AMoRe and MolRep integrated in the CCP4i GUI. These programs simplify the 6-dimensional (6-D) problem of positioning a molecule in the asymmetric unit by running a 3-D rotational grid search followed by a 3-D translational grid search only for the best solution(s) from the rotational search.[27]

However, in many cases the best available search models may have significantly dissimilar regions, and in such cases the task becomes highly laborious and time consuming. Advances aimed at increasing the success rate and throughput fall into two categories: development of new MR algorithms and the encapsulation of MR programs into scripts and GUI's, requiring as little input from the user as possible.

For example, the programs BEAST and its successor PHASER[28] feature two improvements over traditional MR algorithms. They use a maximum likelihood-based (ML) target instead of one based on correlation coefficients as a measure of the quality of the solution and can also use a number of search models simultaneously instead of only one. MR algorithms that perform a simultaneous 6-D search instead of splitting the problem into separate rotational and translational sub-searches can be particularly effective in cases where the search model is highly dissimilar. Alternative search techniques, such as the

evolutionary search algorithm in EPMR,[29] may be used to cope with the 6-D parameter space in a highly efficient manner.

To further streamline the MR procedure, some structural genomic initiatives have incorporated one or more MR programs into their automated software packages. NYSGRC (New York Structural Genomics Research Consortium, www.nysgrc.org) has developed the web-based ASDP (Automated Structure Determination Platform), part of which is an MR server using CNS, AMoRe, MERLOT and MolRep. The TBSGC (Mycobacterium Tuberculosis Structural Genomics Consortium, www.doe-mbi.ucla.edu/TB) have integrated homology model-building systems with the MR program EPMR, simulated-annealing molecular dynamics with CNS, and the bias removal and map reconstruction protocol Shake&Warp. This automated platform generates a number of search models from available databases and processes them as far as the structural model-building/rebuilding step.

It is worth mentioning that initial phases for protein crystal structure solution can be obtained by combining phases from *ab initio* and MR methods.

Model building

The main bottleneck following phase determination is fitting the model to the newly obtained electron density map. Model building has until recently been a particularly slow, tedious and highly manual task involving many hours spent at a computer graphics terminal. Fitting a new structure especially can take many days or even weeks of work, even with good, high-resolution data. There have been tools to aid in chain tracing and fitting, such as BONES, however they are no more than tools, still tied to a manual fitting process. The key to a high-throughput approach is to minimise human input into the process and remove the interactive graphics terminal (and its human operator!) from the loop as far as possible. The past five years or so have seen a great deal of effort in developing algorithms for interpretation of electron density maps and automated model-building tools for protein structures.[30] Programs such as ARP/WARP[31] and RESOLVE,[22] which attempt to automatically interpret the electron density and build in the protein backbone and the side-chains, are now available. The success rates of these programs at present are highly dependent on map quality and resolution (they require moderately high resolution data to fit sidechains automatically) and they often cannot fit all the residues. Thus they do not eliminate the need for model building on the computer graphics terminal, but they can dramatically reduce the amount of manual work that needs to be done. Work is continuing to improve and develop the fitting algorithms to improve the success rate for automated fitting and to lower the resolution required.

The basis of fragment-based screening using X-ray crystallography is to obtain protein–ligand structures by soaking molecular fragments into crystals. Binding of a fragment/ligand can cause localised conformational changes in the protein. One of the challenges for high-throughput screening by X-ray crystallography is to be able to automatically identify and rebuild areas of the protein that differ from the native structure used as the model. The radius of convergence of most refinement programs is too small to cope with such conformational changes, thus the application of the kind of automatic fitting algorithms used in the above programs can be of immense benefit.

The ultimate goal of automation is to put all of these tasks together into a single unified structure determination package that, with a single input step, will locate heavy atoms, generate and improve a phase set and build the model without need for manual input, resulting in a co-ordinate file for the protein. A number of packages, such as AUTOSHARP, SOLVE/RESOLVE and CHART now approach this goal and new packages, such as PHENIX[32] aspire to provide a complete solution within a single environment.

Protein crystallography in drug design

Protein X-ray crystallography has traditionally been viewed by the pharmaceutical industry as a distinctly 'low-throughput' technique; thus its use in drug discovery has been limited to the lead optimisation phase. This involves taking a compound identified as a 'hit' by high-throughput screening (HTS) of the target protein based on a target-specific bio-assay, then making crystals of the complex formed by specific binding of the compound (ligand) to the protein, either by soaking crystals of the protein in a solution of the compound, or by growing co-crystals from a solution containing both the protein and the compound. The crystal structure of the protein–ligand complex is then determined by conventional X-ray crystallographic methods, that is, by fitting a model of the structure to the experimental electron density map. The 3-D atomic model so obtained provides detailed information on the interactions between the protein and the ligand at the atomic level, and so provides an excellent starting point for the synthetic medicinal chemist to begin designing modifications aimed at optimising both the binding affinity and the compound's pharmacokinetic profile.

Even with its traditional shortcomings, protein crystallography has grown in importance within the pharmaceutical industry over the last 10 years. This has largely been fuelled by the availability of many more crystal structures of therapeutic targets. In some pharmaceutical companies, up to a third of all lead optimisation chemistry is guided by information derived from compounds bound to the crystal structure of the target. Several marketed drugs have resulted from these structure-based approaches, perhaps the most notable being the viral enzyme inhibitors such as Agenerase and Viracept for HIV and Relenza for the treatment of influenza. Furthermore, it is estimated that an additional forty compounds generated by structure-based chemistry are currently in clinical trials.[33]

The initial stages in the determination of the crystal structure of a protein–ligand complex parallel those employed in the determination of a *de novo* protein structure by the method of molecular replacement. Additional steps involving the use of difference Fourier methods are then required to determine the structure of the bound ligand. A key step in this process is the search for the precise location of the ligand bound to the protein. Traditionally, crystallographers have achieved this by the use of interactive computer graphics model-building software (typified by the X-LIGAND module[34] in the QUANTA program) to generate orientations and conformers for the ligand until one is found that matches the shape of the experimental electron density. This is usually done by simple visual comparison, though the program does provide semi-automatic tools to perform local searches and to optimise the fit. Fig. 1 illustrates a typical flow-chart for a high-throughput crystallographic ligand-screening experiment.

As a result of the improvements in the speed and efficiency of key technologies discussed in earlier sections, together with complete automation of the process of search and optimisation of the ligand, it has now become feasible to expand the application of protein crystallography beyond the lead optimisation phase. Protein crystallography can now be employed as a much more sensitive screening tool in the lead discovery phase, compared with conventional HTS, as a means to find novel lead compounds that could not otherwise have been discovered.

The application of high-throughput protein crystallography to lead discovery is a strategy that combines features of random

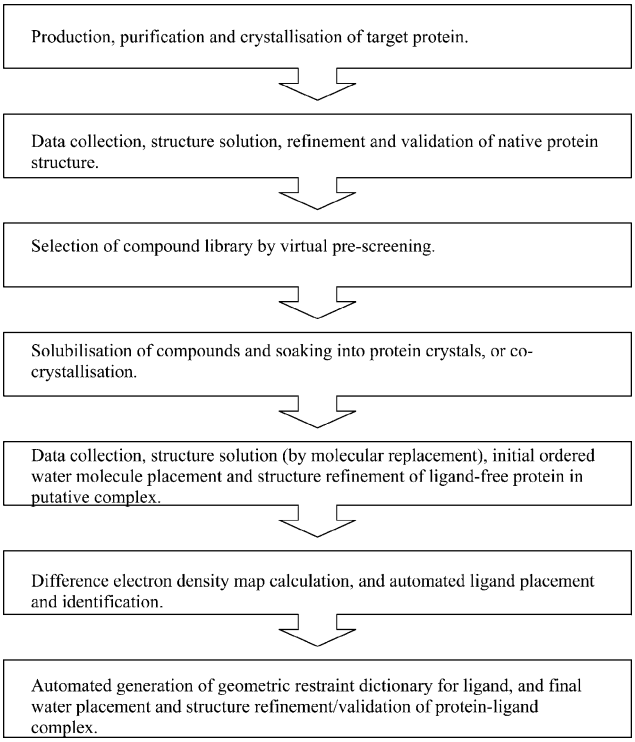

Fig. 1 Typical flow-chart for high-throughput ligand screening experiment.

screening and rational structure-based design.[35] Protein crystallography is more sensitive than the bio-assays used in HTS, typically by a factor of at least 1000 in binding affinity, and also has the advantage, in common with conventional lead optimisation techniques, of furnishing the precise details of the interactions between the protein and the ligand that are needed to progress the fragment molecules through to the structure-based drug design stage. The sensitivity of protein crystallography means that it can detect lower affinity (in the millimolar range) and therefore lower molecular weight bound ligands (known as 'fragments', with molecular weights typically in the range 100–250 Da), compared with HTS (with affinities in the micromolar range and molecular weights of 300–500 Da).

The size of an HTS set has to be much larger than that of a fragment-screening set, simply because the higher molecular weight compounds in an HTS set tend to possess more functional groups than would small fragments, and so there are many more possible combinations of these functional groups to be explored. These low molecular weight fragments are also likely to provide better starting points for lead optimisation, because the more functional groups there are, the greater is the likelihood that one or more of these groups will not interact well with the protein. The fragments can be combined on to a template or be used as the starting point for growing out an inhibitor structure into other pockets on the active site.

In practice, in this 'fragment-based screening' technique, a selection ('library') of fragments is normally soaked, either individually or as mixtures dissolved in a polar organic solvent (usually dimethyl sulfoxide, DMSO, or *N*-methylpyrrolidone, NMP), into the crystals of the target protein, typically for about 1 hour, in order to give the molecules of the compound time to penetrate into the active site of the protein molecules. The concentration of the molecular fragment is typically 50–200 mM. This is a much higher concentration than is used in HTS, and reflects not only the weakness of the interaction being investigated, but also the high concentration of the protein in the crystal ($\sim$ 10 mM). Compounds can be soaked individually or as mixtures (or "cocktails"). If mixtures are used, it is best if the individual compounds are as shape-diverse as possible. The number of compounds that can be soaked as a mixture is limited by the concentration of DMSO that is tolerated by the protein crystals (typically up to 10%), and therefore by the total concentration of organic compound that can be solubilised by the DMSO and by the minimum concentration of an individual compound that is detectable – the latter will clearly depend on the binding affinity to the protein.

Nienaber *et al.*[36] describe a high-throughput method where the target protein crystals are soaked in mixtures of up to 100 shape-diverse molecules at a time that can be distinguished by visual inspection of an interactive computer graphics representation of the difference electron density map. Blundell *et al.*[37] use virtual screening of compounds *in silico* as a pre-screen to identify the most suitable molecular fragments, followed by automated molecular-fragment matching to the difference electron density map and geometry-restrained optimisation using the AutoSolve® software (a component of the HTX® pipeline), in order to rank candidate fragments in cocktails of 4–8 compounds at a time. The set of compounds that go into each cocktail are selected to be as shape-diverse as possible.

As discussed earlier, manual interpretation of the electron density maps is a major bottleneck in the process, because typically a fragment-based screening set will contain about 500 compounds (or about 100 cocktails), and it normally takes even an experienced crystallographer several hours in front of a computer graphics screen to analyse the results for each cocktail. In addition this manual procedure can be very subjective, so that several crystallographers interpreting the same density may disagree as to the precise interpretation, particularly if the resolution of the electron density map is low, and therefore the structural details are not clear. Software such as AutoSolve® is designed to meet the clear need for a fast, completely automatic and totally objective procedure for the structure determination of protein–ligand complexes.

Conclusions

As the drive for increased productivity continues within the pharmaceutical industry, technologies that can improve the success rates in the lead discovery and optimisation process remain a key focus. Over the last decade, protein crystallography has established itself as one of these technologies, with most pharmaceutical companies exploiting structural information in many of their programs. This has been the result of a variety of technological advances, ranging from new methods in molecular biology to novel computer software for analysis of X-ray data, that now mean crystal structures of therapeutic targets can be determined in a more timely and resource-efficient manner. The ability to rapidly and routinely obtain structures of lead compounds bound to these drug targets has also resulted in a significant increase of structure-based design in many lead optimisation programs. Indeed, high-throughput crystallography is now being utilised in a novel approach for lead discovery in which fragment libraries are screened to identify new chemical entities that can be optimised into drug candidates. It may be that protein crystallography is finally fulfilling its true potential within the drug discovery process.

References

1 International Human Genome Sequencing Consortium, *Nature (London)*, 2001, **409**, 860–921.
2 S. F. Campbell, *Clin. Sci.*, 2000, **99**, 255–260.
3 S. K. Burley, *Nat. Struct. Biol. Suppl.*, 2000, 932–934.
4 U. Heinemann, K. Bussow, U. Mueller and P. Umbach, *Acc. Chem. Res.*, 2003, **36**, 157–163.

5 S. P. Chambers, *Drug Discovery Today*, 2002, **7**, 759–765.
6 R. F. Service, *Science (Washington, D. C.)*, 2002, **298**, 948–950.
7 R. C. Stevens, *Structure (London)*, 2000, **8**, R177–R185.
8 A. M. Edwards, C. H. Arrowsmith, R. Hui, F. Marino, K. Yamazaki, A. Savchenko and A. Yee, in *Protein Structure: Determination, Analysis, and Applications for Drug Discovery*, ed. D. I. Chasman, Marcel Dekker Inc., New York, 2003.
9 A. McPherson, *Crystallization of Biological Macromolecules*, Cold Spring Harbor Laboratory Press, New York, 1999.
10 J. R. Luft, R. J. Collins, N. A. Fehrman, A. M. Lauricella, C. K. Veatch and G. T. DeTitta, *J. Struct. Biol.*, 2003, **142**, 170–179.
11 G. Sulzenbacher, A. Gruez, V. Roig-Zamboni, S. Spinelli, C. Valencia, F. Pagot, R. Vincentelli, C. Bignon, A. Salomoni, S. Grisel, S. Maurin, C. Huyghe, K. Johansson, A. Grassick, A. Roussel, Y. Bourne, A. Perrier, L. Miallau, P. Cantau, E. Blanc, M. Genevois, A. Grossi, A. Zenatti, V. Campanacci and C. Cambillau, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2002, **58**, 2109–2115.
12 B. Rupp, B. W. Segelke, H. I. Krupka, T. P. Lekin, J. Schafer, A. Zemla, D. Toppani, G. Snell and T. Earnest, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2002, **58**, 1514–1518.
13 E. R. Bodenstaff, F. J. Hoedemaeker, M. E. Kuil, H. P. M. de Vrind and J. P. Abrahams, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2002, **58**, 1901–1906.
14 C. L. Hansen, E. Skordalakes, J. M. Berger and S. R. Quake, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 16531–16536.
15 M. P. Blakely, M. Cianci, J. R. Helliwell and P. J. Rizkallah, *Chem. Soc. Rev.*, 2004, this issue (DOI: 10.1039/b312779f).
16 A. Sharff and H. Jhoti, *Curr. Opin. Chem. Biol.*, 2003, **7**(3), 340–345.
17 S. W. Muchmore, J. Olson, R. Jones, J. Pan, M. Blum, J. Greer, S. M. Merrick, P. Magdalinos and V. L. Nienaber, *Structure (London)*, 2000, **8**, R243–R246.
18 C. Yang, J. W. Pflugrath, D. A. Courville, C. N. Stence and J. D. Ferrara, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2003, **59**, 1943–1957.
19 T. M. McPhillips, S. E. McPhillips, H.-J. Chiu, A. E. Cohen, A. M. Deacon, P. J. Ellis, E. Garman, A. Gonzalez, N. K. Sauter, R. P. Phizackerley, S. M. Soltis and P. Kuhn, *J. Synchrotron Radiat.*, 2002, **9**, 401–406.
20 A. G. W. Leslie, H. R. Powell, G. Winter, O. Svensson, D. Spruce, S. McSweeney, D. Love, S. Kinder, E. Duke and C. Nave, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2002, **58**, 1924–1928.
21 V. S. Lamzin and A. Perrakis, *Nat. Struct. Biol.*, 2000, **7**, 978–981.
22 T. C. Terwilliger, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2002, **58**, 1937–1940.
23 P. D. Adams, R. W. Grosse-Kunstleve, L.-W. Hung, T. R. Ioerger, A. J. McCoy, N. W. Moriarty, R. J. Read, J. C. Sacchettini, N. K. Sauter and T. C. Terwilliger, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2002, **58**, 1948–1954.
24 J. S. Brunzelle, P. Shafaee, X. Yang, S. Weigand, Z. Ren and W. F. Anderson, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2003, **59**, 1138–1144.
25 Z. Dauter, *Curr. Opin. Struct. Biol.*, 2002, **12**(5), 674–678.
26 G. Taylor, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2003, **59**, 1881–1890.
27 R. W. Grosse-Kunstleve and P. D. Adams, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2001, **57**, 1390–1396.
28 R. J. Read, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2001, **57**, 1373–1382.
29 C. R. Kissinger, D. K. Gehlhaar and D. B. Fogel, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 1999, **55**, 484–491.
30 T. J. Oldfield, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2003, **59**, 483–491.
31 A. Perrakis, R. Morris and V. S. Lamzin, *Nat. Struct. Biol.*, 1999, **6**, 458–463.
32 P. D. Adams and R. W. Grosse-Kunstleve, *Curr. Opin. Struct. Biol.*, 2000, **10**(5), 564–568.
33 L. W. Hardy and A. Malikayil, *Curr. Drug Discovery*, Dec. 2003, 15–20.
34 T. J. Oldfield, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2001, **57**, 696–705.
35 C. L. M. J. Verlinde, H. Kim, B. E. Bernstein, S. C. Mande and W. G. J. Hol, in *Structure-based Drug Design*, ed. P. Veerapandian, Marcel Dekker, New York, 1997, pp. 365–394.
36 V. L. Nienaber, P. L. Richardson, V. Klighofer, J. J. Bouska, V. L. Giranda and J. Greer, *Nat. Biotechnol.*, 2000, **18**, 1105–1108.
37 T. L. Blundell, C. Abell, A. Cleasby, M. J. Hartshorn, I. J. Tickle, E. Parasini and H. Jhoti, in *Drug Design: Special Publication*, ed. D. R. Flower, Royal Society of Chemistry, Cambridge, 2002, **vol. 279**, pp. 53–59.