J Struct Funct Genomics (2010) 11:211–221
DOI 10.1007/s10969-010-9092-9

To automate or not to automate: this is the question

M. Cymborowski · M. Klimecka · M. Chruszcz · M. D. Zimmerman · I. A. Shumilin · D. Borek · K. Lazarski · A. Joachimiak · Z. Otwinowski · W. Anderson · W. Minor

Received: 23 December 2009 / Accepted: 14 May 2010 / Published online: 6 June 2010

Abstract New protocols and instrumentation significantly boost the outcome of structural biology, which has resulted in significant growth in the number of deposited Protein Data Bank structures. However, even an enormous increase of the productivity of a single step of the structure determination process may not significantly shorten the time between clone and deposition or publication. For example, in a medium size laboratory equipped with the LabDB and HKL-3000 systems, we show that automation of some (and integration of all) steps of the X-ray structure determination pathway is critical for laboratory productivity. Moreover, we show that the lag period after which the impact of a technology change is observed is longer than expected.

Keywords Automation · Databases · Data collection · Ligand screening · Structural genomics · Structure determination

M. Cymborowski · M. Klimecka · M. Chruszcz · M. D. Zimmerman · I. A. Shumilin · W. Minor (✉)
Department of Molecular Physiology and Biological Physics, University of Virginia, 1340 Jefferson Park Avenue, Charlottesville, VA 22908, USA
e-mail: wladek@iwonka.med.virginia.edu

D. Borek · Z. Otwinowski
Department of Biochemistry, University of Texas, Southwestern Medical Center in Dallas, 5323 Harry Hines Blvd, Dallas, TX 75390-8816, USA

K. Lazarski · A. Joachimiak
Structural Biology Center, Bioscience Division, Argonne National Laboratory, Argonne, IL 60439, USA

W. Anderson
Molecular Pharmacology & Biological Chemistry, Northwestern University, 303 E. Chicago, Chicago, IL 60611-3008, USA

M. Cymborowski · M. Klimecka · M. Chruszcz · M. D. Zimmerman · I. A. Shumilin · D. Borek · A. Joachimiak · Z. Otwinowski · W. Anderson · W. Minor
Midwest Center for Structural Genomics
URL: http://www.mcsg.anl.gov

M. Cymborowski · M. Klimecka · M. Chruszcz · M. D. Zimmerman · I. A. Shumilin · D. Borek · A. Joachimiak · Z. Otwinowski · W. Anderson · W. Minor
Center for Structural Genomics of Infectious Diseases
URL: http://www.csgid.org

Abbreviations

ALS	Advanced Light Source
CDP	Cytidine-5′-diphosphate
CSGID	Center of Structural Genomics for Infectious Diseases
MCSG	Midwest Center for Structural Genomics
NSLS	National Synchrotron Light Source
PDB	Protein Data Bank
SG	Structural genomics
SGC	Structural Genomics Consortium

Introduction

During last 10 years, several high throughput—and even high output—structure determination pipelines (mostly using X-ray diffraction methods) were developed by a number of multi-institutional consortia. They all share the same goal: rapid progress from the cloning of a protein gene to the determination and deposition of its structure into the Protein Data Bank (PDB) [1].

The most productive X-ray crystallography pipelines established by some structural genomics (SG) groups are capable of depositing 200 structures per year. This rate of structure determination would not be possible without the substantial effort that these groups put into optimization and automation of all stages of the structure determination process: cloning, expression, purification, crystallization, data collection, processing, phasing, model building, structure refinement, validation and deposition. While the whole process is not yet fully automated, both hardware and software tools and protocols have been developed to partially or fully automate nearly every stage of the process.

In contradiction to anecdotal experience, it has been shown that there is no clear single bottleneck in the structure determination process [2], except perhaps at the point when it is necessary to engage the brain of the researcher. The most productive SG centers developed significant automation of the structure elucidation process. In many cases in this automated environment, the first time when the researcher's brain is fully engaged is the biological interpretation of the 3-D protein structure, i.e. the process of analysis of data, integrating results and writing the publication.

As the analysis and description of the relationship between a protein's structure and function has not yet been automated, the most successful SG groups publish only a fraction of their structures in peer-reviewed journals. However, high-impact research has been published by SG groups. An analysis of PDB data shows that out of 6,955 structures reported by SG centers around the world since 2005, 3.7% of those structures were reported in the high-impact journals *Nature*, *Science*, *Cell* and *PNAS*.

We describe our automation protocols to improve the efficiency of various steps of the high-throughput structure determination pipeline, as a part of our work in both the Center of Structural Genomics of Infectious Diseases (CSGID) and the Midwest Center for Structural Genomics (MCSG). In particular, we present and discuss automation and protocols that are applicable to a small- or medium-sized laboratory, such as the one at University of Virginia. The majority of structures solved by these consortia were determined by means of X-ray crystallography, though other methods such as NMR may be used in high-throughput structure determination. As our automation experience deals almost exclusively with the X-ray crystallography pipeline, we will focus on this technique in this work.

Automation of cloning, expression, purification and crystallization

It is evident that the production of diffraction-quality macromolecular crystals is the most challenging and expensive step in the process that leads to determination of the macromolecular structure. The four steps are tightly linked to one another and should be treated as a single process leading to high-quality crystals. Seldom are expressed recombinant proteins soluble, purification straightforward, and diffraction-quality crystals obtained with only initial screening. In reality, each step has to be performed several times. For example, the Structural Genomics Consortium (SGC) reports the use (on average) of $\sim$20 different constructs for each single successful structure determination process [3].

In practice, many automation tools have been developed for virtually all steps of the protein production and crystallization pipeline. For example, cloning and expression has been automated using technologies such as ligation-independent cloning methods [4] and Biomek/Multimek (Beckman Coulter) liquid handling systems [5]. Protein purification has been automated by a number of groups using high-capacity liquid chromatography systems, such as the AKTA Express (GE Healthcare) series of instruments, e.g. [6–8].

As (arguably) the most labor-intensive part of the process, high-throughput crystallization and crystal observation has spawned many automation technologies, most of which are commercially available. These include liquid handling systems for preparation of custom crystallization formulations like the Alchemist II (Rigaku), Biomek (Beckman-Coulter), Matrix Maker (Emerald BioSystems), and Freedom EVO (Tecan); plate setup robots like the Phoenix RE (Art Robbins/Rigaku) and Mosquito (TTP LabTech); plate observation systems like the Minstrel (Rigaku), CrystalFarm (Bruker), and CrystalPro (Tritek); and even crystal harvesting robots such as the Crystal Harvester (Bruker) [9, 10].

The most difficult part of this process is the connection of these disparate components into an integrated workflow [11]. In addition, the degree various steps in the protein production and crystallization pipeline should be partially or fully automated depends on an analysis of the bottlenecks, impact and cost. Some steps, especially the preparation and observation of crystallization plates almost always demands automation of some kind. However, in small or moderate-size operations, such as a single-principal investigator laboratory, fully automated cloning and expression methods are not necessary, as more traditional methods of expression (e.g., in regular 1–3 L fermentation flasks) may be used. However, some way of integrating and managing data from a blend of manual and automated approaches is necessary.

The LabDB system is the central database which tracks cloning, expression, purification, and crystallization experiments in our laboratory. A schematic of the LabDB system is presented in Fig. 1. LabDB is designed to input information

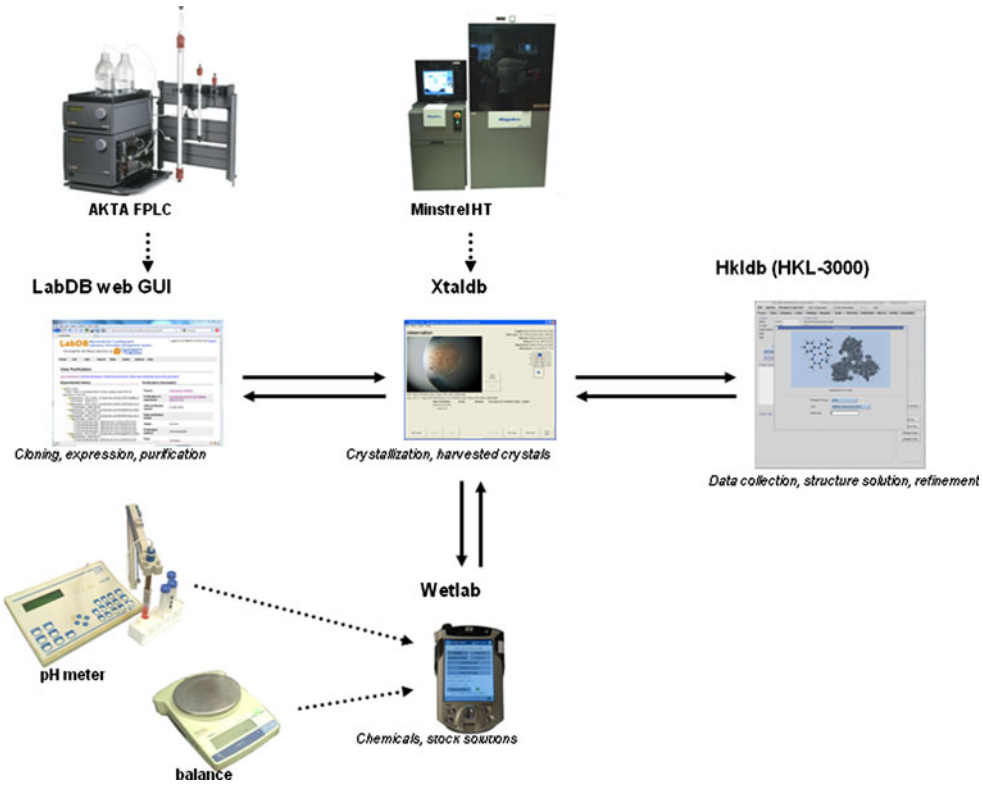

Fig. 1 A schematic overview of the LabDB system. The four components of the system are shown connected by *solid arrows*, while laboratory systems from which data are automatically extracted are shown connected by *dotted lines*

both from manual entry and from automated systems. The manual components are the PHP-based web interface for LabDB and the Xtaldb system [12].

The cloning and expression pipeline in the laboratory is largely not automated, and thus data for these types of experiments are entered mostly by hand into the database. However, one chromatography step of the purification process is integrated into the LabDB system by a custom module that imports information directly from the AKTA's UNICORN software system. Every time a chromatographic separation is executed, detailed information about the process, including the chromatogram, peak heights, etc. is imported into the system.

The Xtaldb [12] component of LabDB contains an interface for semiautomatically adding images and annotating crystallization drops. Recently, we have developed a module to automatically import into LabDB images and drop annotations made automatically by a Minstrel HT (Rigaku, Inc) system, by communicating with the CrystalTrak database (Rigaku, Inc.). The laboratory in Virginia also makes use of a Mosquito dispensing robot (TTP Labtech, LLC.) and other tools like multichannel pipetters to semiautomatically generate initial 96-well crystallization plates. However, automation does introduce serious limitations, as only crystallization plates compatible with the robots—typically those with the standard 96-well Society for Biomolecular Sciences (SBS) microplate footprint—may be used.

The process of crystal growth optimization in our laboratory, on the other hand, is largely done manually, in 24-well plates. A number of large, high-throughput SG labs also tend to use automatic processes for setting up initial screens of protein but the process of optimization is largely done by hand. A fair amount of effort has been put into generating customized crystallization screens, identifying optimized conditions proven successful for several other proteins. Some of the optimized screens have subsequently been commercialized and have joined the ranks of more "traditional" sets of screens e.g., the JCSG Core [13], and JCSG + [14] screens (QIAGEN, Inc.).

Crystallization optimization is also difficult to automate because often changes need to be made to the recalcitrant protein itself in order to get it to crystallize. Some protocols have been developed to approach this problem in at least a semiautomatic way: e.g. limited proteolysis [3, 15], protein methylation [16, 17] (see also Fan & Joachimiak, this issue), automated domain design (Babnigg and Joachimiak, this issue; and http://bioinformatics.anl.gov/cgi-bin/tools/pdpredictor) and additive screening [18]. The latter procedure, in addition to potentially enhancing the crystallization of a target protein, can also provide important information about the function and/or mechanism of a protein in the absence of other biochemical information. Other approaches require returning to a prior step in the pipeline, such as the generation of alternative constructs including e.g. surface entropy reduction [19].

The Wetlab component of the LabDB system does little to automate the actual work of the production of stock solutions of chemicals, but it does automate, with a much greater level of detail, the record-keeping associated with chemical stocks and stock solutions. Every bottle of a reagent is entered into a database, and by using a balance, pH meter, and barcode printer connected by a port server directly to the LabDB database, the process of labeling and recording prepared stock solutions is automated. This is especially advantageous in that every stock solution is annotated with information that is not ordinarily included whenever such a solution is prepared by hand: e.g., the lot number of the chemical, the date it was received, etc.

Information in LabDB about the cloning, expression, purification, and crystallization of CSGID and MCSG projects are transferred to the central databases for these SG efforts by means of XML files. The XML formats, which list details of each kind of experiment, are specified by means of XML Schema documents. These specifications may be used to validate a given XML file for syntactic correctness. An automated script queries the LabDB database, generates the file and places it in a publicly accessible location. These files are then downloaded regularly by the CSGID and MCSG databases, parsed, and their experimental information is imported.

The choice of the best path (or protocol) for navigating the protein production and crystallization pathway, particularly given that information must be integrated from both manual and automatic sources, is a difficult one that consumes time and money.

Data collection and structure solution

Currently there are over 125 synchrotron stations in the world that are suitable for (and many are dedicated for) X-ray macromolecular diffraction experiments. In recent years, over 80% of PDB deposits report the use of a synchrotron source for diffraction experiments [20]. In comparison with experiments performed only 20 years ago, even the simplest synchrotron stations are highly automated.

The automation was possible thanks to both hardware and software developments at the beamlines. Software development is especially important in integration of different hardware components and enhancement of the researcher's ability to control the diffraction experiment. In most cases the experimenter uses a single, usually user-friendly interface [21]. Moreover, such software allows for remote data collection, and thanks to this software, so-called 'mail-in' crystallography is becoming more popular [22, 23]. The ability to collect data from distant locations via remote access to synchrotron beamlines would be not possible without development of robotic systems for storing and mounting of crystals [24]. Thus a lot of effort has been put into development of such systems [25, 26]. Currently automatic mounting systems are available from commercial suppliers and in many cases are standard additions to home diffractometers. The solution of problems connected with crystal mounting and centering leads to the development of fully automated beamlines and diffractometers which are very useful for extensive crystallographic screening of potential small-molecule ligands.

Although fully automated systems are capable of high-throughput crystal mounting and data collection, their application does not necessarily instantly and substantially impact the productivity of a synchrotron station (Fig. 2). Even though the diffraction experiment seems to be relatively simple, the fact that the experimenter has limited control of crystal quality can make automation of data collection very challenging. Problems with automation start at the beginning of the diffraction experiment: namely, with crystal centering. Centering is not always simple even for humans, therefore the process may be quite difficult to automate, especially for samples which are suboptimally cryocooled or if the crystals are very small [27]. In such cases, centering may require the use of X-ray or UV radiation [28, 29]. Once the crystal is well centered, its quality has to be evaluated and scored. This step of the data collection process is the most critical, as the strategy of the best diffraction intensity recording protocol is based on initial diffraction images. The correct determination of crystal symmetry, unit cell parameters, mosaicity, and estimation of the crystal's survival time in an X-ray beam with a particular wavelength and intensity, all very strongly affect the quality of the data, and therefore the quality of the subsequent crystal structure. Moreover, proper measurement of the strongest intensities is very important for the choice of strategy during collection of data used for structure solution in both MR and SAD/MAD methods [30].

Successful structure solution and refinement are the best validation methods of the data collection process. Having this in mind, automatic or semiautomatic systems used during the structure determination process should be designed to provide the best possible structures, not just the best data sets. So, ideally the structure solution process should be done in parallel with data processing. HKL-3000 [31], through the integration of data collection, processing, structure solution and refinement, provides the researcher in many cases with an initial model of the structure, when the crystal is still on the goniostat. The HKL-3000 pipeline incorporates many formidable and widely used programs like CCP4 [32], SOLVE [33], RESOLVE [34], MLPHARE [35], SHELXD and SHELXE [36], ARP/wARP [37], DM [38], MOLREP [39], REFMAC [40], and Molprobity [41].

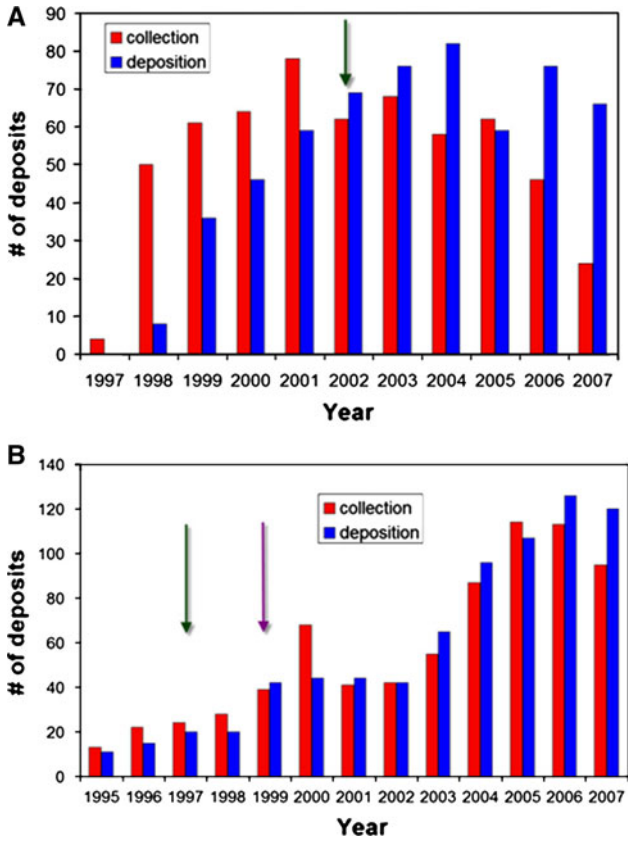

Fig. 2 **A** Number of data sets collected and structures deposited for beamline 5.0.2 at ALS. The *green arrow* indicates the introduction of an automatic crystal mounting system. **B** Number of datasets collected and structures deposited for beamline X4A at NSLS. The *green arrow* indicates introduction of an automatic image plate detector, and the *purple line* marks introduction of a CCD detector

The system, when run in semiautomatic mode, provides the experimenter the ability to check the most important parameters defining the quality of the diffraction data and gives insight into the particular steps of the structure elucidation process, which is divided into six steps [31]. In most cases, using the default settings of the program through these six steps (the "six click" approach) in HKL-3000 results in a highly complete model of a macromolecule. Moreover, such a semiautomatic pipeline of structure determination at every step provides feedback to the experimenter and in the worst case shows why a particular experiment failed, which is not possible in the case of "brute force" automation.

Counterintuitively, the failure of the system to generate a satisfactory model of a given structure is very beneficial for developmental work on difficult cases that cannot be solved by standard approaches, and often leads to improvement of the semi-automated algorithms [42]. When a stubborn structure is successfully built and refined, it is used a posteriori as a test case. Specifically, parameters of the structure solution are varied, and the settings that produce the "best" initial electron density map (as measured by correlation of the map to the final refined model) are incorporated into the structure solution pipeline.

Sometimes projects that seem to be very easy may in fact turn out to be very challenging. Despite having a data set of reasonable quality and properly determined space group, the structure solution or model building occasionally fails for unknown reasons. Our experience shows that most cases are caused by mislabeling of the crystals (or even the proteins used for the crystallization) and as a result the wrong sequence is provided to the structure determination pipeline, causing an unnecessary waste of time. In similar cases the "brute force" approach may be the only way to successfully overcome problem due to an incorrect polypeptide sequence. For example, systems like BALBES [43] or MrBUMP [44] may efficiently try many different models [45] and hopefully return a proper solution. In truly "hopeless" cases, the use of thousands of models may be necessary [46]. However, a simple check of the unit cell parameters and their comparison with unit cell parameters reported in the PDB may immediately show that instead of a Nobel-prize-winning molecule, the experimenter may have crystallized glutathione S-transferase (GST), or some other well known component of the expression system. If the map is of sufficiently high resolution and quality, and the automatic model building algorithm builds the polypeptide backbone but unexpectedly fails to assign sequence, it may be possible to use the density itself to sequence a fragment of the protein and use that fragment to search for the correct protein in sequence databases.

The data collection is the last experimental step in the crystal structure determination process, and errors made at this step may nullify successful work from several prior steps. Therefore it is worth immediately checking the results of this process. In our practice (3–4 "synchrotron trips" per year, ~1,200 crystals screened, and ~600 datasets collected), we noticed that in order to maximize productivity of data collection it is worth taking some crystals still present in crystallization plates. Usually we begin data collection using pre-cryocooled crystals and immediately proceed to structure solution, if possible. Dependent on the initial results, the next crystals from the same project will be screened to search for a higher resolution dataset, but if one is not found, we are able to search for improved soaks conditions while still at the synchrotron. In this process HKL-3000 plays a central role. HKL-3000, in connection with HKLdb, contains all information necessary for efficient data collection and structure solution. Information on the crystals which are placed in the X-ray beam are linked to the crystallization database (Xtaldb), from which one may retrieve information on compounds used for soaking experiments (for example).

A fast structure solution and refinement protocol provides electron density maps rapidly, which may be used to determine if the structure contains a bound small molecular agent. This approach leaves time to concentrate on stubborn projects, as less difficult ones are quickly classified as "solvable" and "refineable."

Ligand screening and identification

In many cases, the structures of the *apo*-forms of proteins provide limited information about protein mechanism of action. For that reason, most biologists, biochemists and drug developers are interested ultimately in gaining insights into interactions of the protein with ligands and effectors related to the protein function or the regulation thereof. The direct approach to obtain this information is the determination and analysis of the *holo*-forms of protein structures. Apart from a few lucky cases where bound ligand is retained by the protein throughout purification and crystallization, there are two ways to obtain crystals of protein–ligand complexes: cocrystallization and soaking.

The rationale for cocrystallization is that the protein is more likely to bind a ligand in solution than in a crystal where the packing and crystallization interactions might limit or obstruct the formation of the protein–ligand complex. Cocrystallization is certainly the technique of choice when the interacting ligand is a macromolecule that is not able to penetrate the crystal for sterical reasons or when ligand binding induces large protein conformational changes. It is also routinely used when only a few interacting partners are studied. However, cocrystallization is poorly suited for extensive ligand–protein binding studies. In many instances, the presence of an additional component in the crystallization solution alters the conditions at which crystals form even when ligand does not bind to the protein in an ordered way. This effect worsens as the concentration of the added component increases. It necessitates searching for a new optimal crystallization condition, making the overall study much more resource-intensive.

Soaking implies that the ligand diffuses into pregrown and, possibly, stabilized crystals of the *apo*-form of the protein and interacts with the binding site, which is not constrained by crystal contacts. By soaking protein crystals in cocktails of multiple ligands, more compounds may be screened with the same number of crystals, further increasing the throughput. Indeed, crystallographic screening of chemical libraries is now employed in fragment-based lead discovery in drug design for detecting the binding of low affinity, low molecular weight compounds [47]. The screening is commonly done with soaks containing from four to ten compounds in millimolar concentrations, which enable binding of even low affinity ligands. The desirable outcome of the cocktail soak screening is the presence of additional electron density followed by direct identification of the bound ligand from the soaked structure. The chances of correct ligand identification increase as the crystals diffract to higher resolution, the ligand occupancy is higher, and as the different ligands present in the cocktail differ significantly in molecular shape or possess distinct functional groups that can be identified through interaction with the protein. Cocktail components can be more directly detected and identified in a structure by the introduction of atoms providing an anomalous signal.

The throughput of crystal screening can be significantly increased with automation. Cocktail approach requires the collection of many more datasets on many more crystals than in the primary structure determination process. Accordingly, the use of synchrotron radiation and robot-assisted mounting becomes even more important. The solution of the soaked structures is usually straightforward since the space group and unit cell dimensions are typically very similar to that of the *apo*-form of the crystal. In HKL-3000, data processing is followed by map generation using phases from the model of the native protein, or if there are changes in the crystal form, by full molecular replacement (MR) by MOLREP [39]. This is coupled with a module that semiautomatically analyzes the fit of the soak components into any unexplained electron density regions using the predefined set of the cocktail component structures (Fig. 3a). The analysis, which is done by RESOLVE [48], produces a set of ligand structures ranked according to their fit to the density [49]. Figure 3 illustrates the application of the ligand analysis module to the identification of an unidentified component bound to the structure of APC7551, a universal stress protein from *Archaeoglobus fulgidus*, which was soaked in a cocktail of ten compounds. The module properly identified the bound ligand as cytidine-5′-diphosphate (CDP) (Fig. 3b), which has a better correlation with the electron density than the other components of the soak, such as trehalose (Fig. 3c). The scores showing the quality of the fit of each compound to the unknown density are shown in Fig. 3d. Other compounds have very similar quality of fit scores to CDP, such as cytidine-5′-monophosphate (CMP) and cytidine, which is to be expected given their chemical similarity. In this case, determining which compound has the best fit (CDP, CMP or cytidine) requires manual visual inspection of the ligand models. (A different cocktail design better suited for uniquely identifying the best binding compound would have used a set of more dissimilar reagents.) Our experience shows that this human intervention is almost always required, because ligand assignment can be impeded by conformational changes in the protein or partial disorder in the ligand structure.

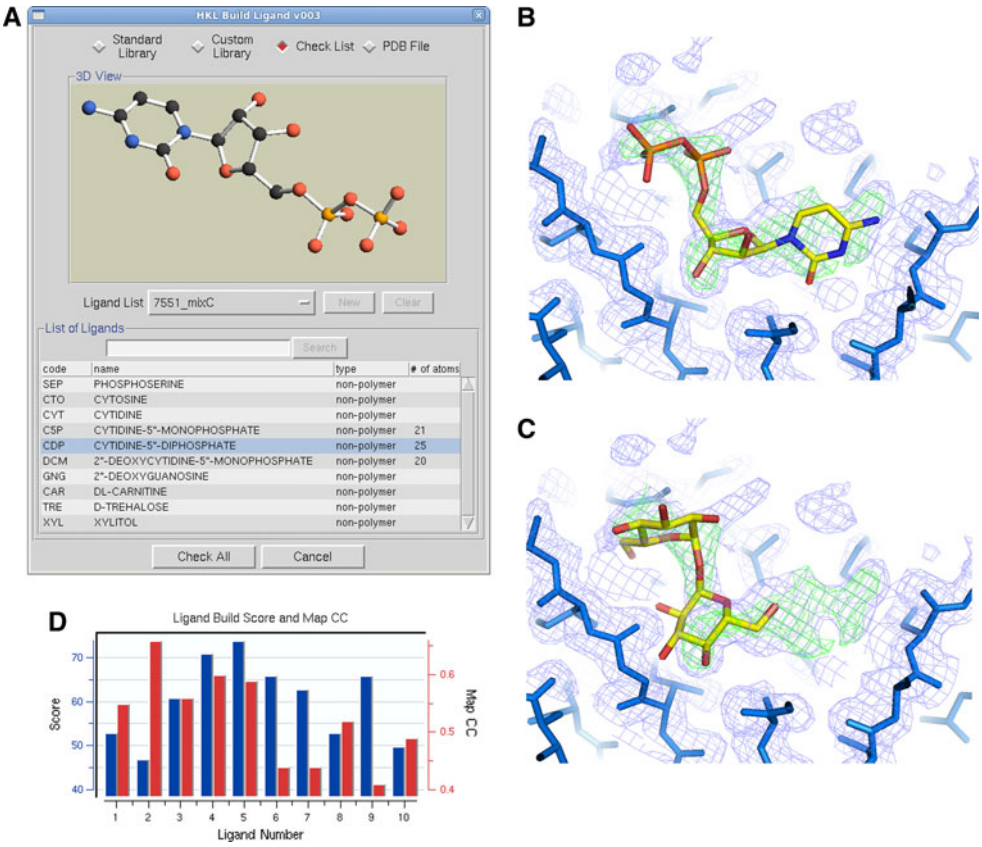

Fig. 3 Application of the ligand analysis module in HKL-3000 to the study of ligand binding properties of APC7551, a universal stress protein from *Archaeoglobus fulgidus*. **A** The module interface that describes the soak composition. **B** The bound component of the soak cocktail, CDP, is automatically fit into the additional electron density ($2F_o - F_c$ is *blue*, $F_o - F_c$ is *green*). **C** A fit of an incorrect component of the soak cocktail, trehalose, into the same density. **D** A scoring diagram that describes the fit of the soak components into electron density (CDP is number 5, trehalose is number 9)

Refinement, validation and deposition

The modern refinement process is highly automated, especially for structures determined at resolution 2.5 Å or better. However, manual inspection of the map should be a compulsory practice for every protein structure ready for deposition. Parameters like R, R_{free}, clashscore [41], and Ramachandran plot statistics describe only the global correctness and quality of the structure. Flexible parts of the protein can be identified by high values of the displacement parameter (B-factor), and usually require manual correction. Small errors in mobile parts of the protein may not significantly affect global statistics but may be important for interpretation of the structure–function relationship. Similarly, the use of automatic procedures for identification and refinement of ligands requires manual inspection and verification even for relatively high resolution structures.

Automatic procedures quite often fail to properly identify and refine metal ions, but their proper identification is very important. Around 20% of all PDB deposits report the presence of ordered metal ions adjacent to sites important for the biological activity of the macromolecule. Analysis of the PDB shows that for medium resolution data (2.0–2.5 Å), the environments of many zinc ions are not identified or refined properly (Fig. 4). In many cases, identification and/or refinement of the metal binding sites are clearly incorrect [50], when compared to very high resolution structures in the Cambridge Structural Database [51]. PDB deposits do not contain any information about the procedures that were used to identify and refine metal ions, but rarely are anomalous data from an additional experiment at the appropriate wavelength used to identify possible metal ions unambiguously. Similarly, the drive to automate the process of electron density map interpretation has increased the number of deposits that contain unidentified small molecule agents. The fraction of structures with clearly marked unknown ligands is higher for higher resolution structures [52], as it is more difficult to place an erroneous arbitrary ligand into a high resolution map.

Regardless of the degree of automation, the final structure quality should be carefully assessed by a human being. What quality statistics we should expect? When a structure is refined with HKL-3000, the experimenter can, at any point of the refinement procedure, compare the statistics of the current refinement with the average R and geometry statistics derived from recent PDB deposits in the same resolution range [53, 54]. The HKL-3000 refinement module also shows structure quality guidelines agreed between the NIAID Infectious Diseases SG centers, namely the Seattle Structural Genomics Center for Infectious Disease (SSGCID) [55] and the CSGID [56]. These guidelines mandate that the structure meet quality criteria

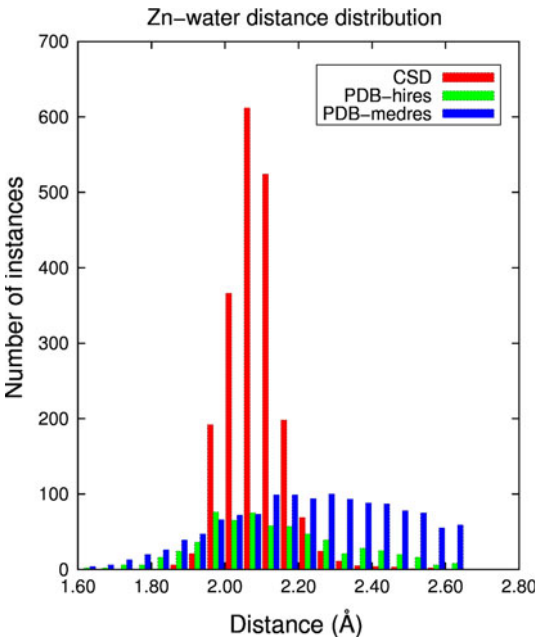

Fig. 4 Distributions of Zn-to-water-oxygen distances in the Cambridge Structural Database (CSD, *red*), a set of high resolution PDB structures (<1.5 Å), and a set of medium resolution PDB structures (2.0–2.5 Å)

stricter than those derived from recent PDB structures (http://www.csgid.org/csgid/cake/pages/sg_metrics). HKL-3000 also uses Molprobity for validation of model geometry, and in addition, a tool is provided that can remove excessive waters based on their B-factor distribution compared to the average B-factor for the structure under refinement. The expected number of waters for structures of similar size and resolution can also be used as a reference (Fig. 5).

The handling of water molecules in HKL-3000 is an example of an automatic procedure well integrated into the semiautomatic refinement process performed with Refmac [40] and its use can sometimes significantly improve the structure quality. Sometimes, the reprocessing of raw diffraction data and rerefinement of already deposited structures not only improves refinement statistics, but may lead to better identification of structural details [57] and more complete models (Fig. 6). In an ideal world, a significantly improved model should be redeposited in the PDB. Although most software developers implement significant improvement of refinement procedures, the number of redeposited structures is below 2% of newly deposited structures, as shown by analysis of the PDB.

There is one process that seems to be easy to automate, but several attempts to fully automate the process of deposition and validation of protein models and crystallographic experimental parameters have failed, including one attempted by MCSG. There is a rising gap between the growing number of protein models and the ability to process and analyze the resulting data in a complex way. The attempt to analyze even simple fields in the PDB header, such as the temperature of crystallization, shows that existing data are not fully reliable [58]. The recent retraction of 11 fraudulent PDB deposits [59] shows the necessity for uniform validation of protein models and uniform validation of experiments that lead to structure solution. Also needed is wider accessibility of raw data in the form of diffraction images—at the moment, only certain groups like the CSGID and the Joint Center for Structural Genomics (JCSG) have a policy of making diffraction images publicly available. The lack of raw data and uniform validation tools makes global analysis of the PDB very difficult, as the creation of database from data submitted into PDB requires curation and editing of impossible values and resolution of many inconsistencies. The creation of uniform, automatic validation tools would simplify the work of journal referees.

Conclusions

It is obvious that automation of any single step of the structure determination pipeline is capable of saving a significant amount of time for the experimenter and enabling the process to be run in a high-throughput manner. However, the advantage conferred by automation of a single step in the multi-step process may be greatly diminished if the automated step is not tightly integrated with other parts of the structure determination pipeline. A series of incremental improvements generates a multiplicative gain in efficiency; so high output is the result of overall efficiency rather than very high efficiency of one individual step.

The use of a single general protocol for cloning, expression, purification and crystallization of challenging proteins will leave too many structures unsolved [52]. Likewise, an opposite approach, namely the development and application of a customized and separate protocol for every single protein, makes both high throughput and high output impossible goals. Any automation approach must have high throughput (output), but be flexible enough to handle multiple protocols.

There are several excellent systems that handle almost whole process from crystal to deposit like PHENIX [60], AUTOSHARP [61], SOLVE/RESOLVE [48], ANTS [62], ELVES [63], CCP4 [32] or AUTORICKSHAW [64]. In our (admittedly biased) opinion, HKL-3000 is the most complete system, integrated with crystallization, protein production and other relevant databases. Analysis of the PDB shows that HKL-3000 was used in the solution of over 1,000 PDB structures, solved both by SAD/MAD and MR techniques, which indicates its robustness.

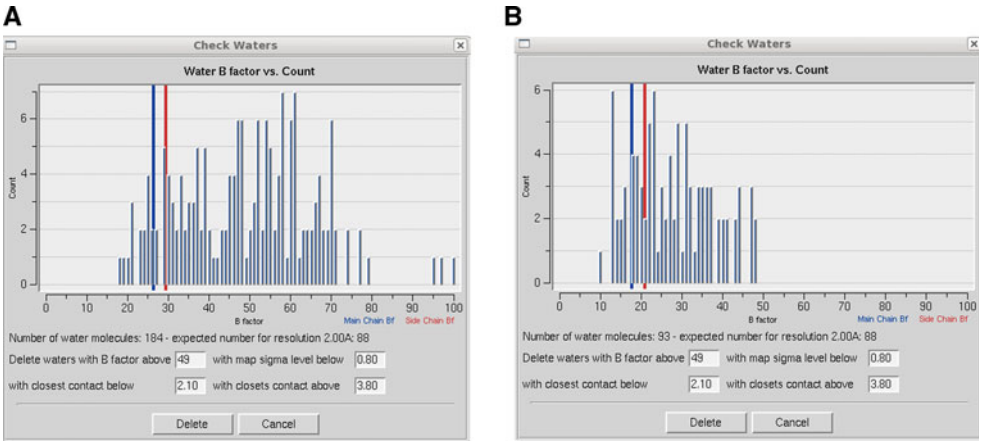

Fig. 5 The water molecule validation window in HKL-3000. **A** The original distribution of water oxygen atoms by B-factor in PDB deposit 3EME. **B** The distribution of water oxygen atoms by B-factor after structure rerefinement

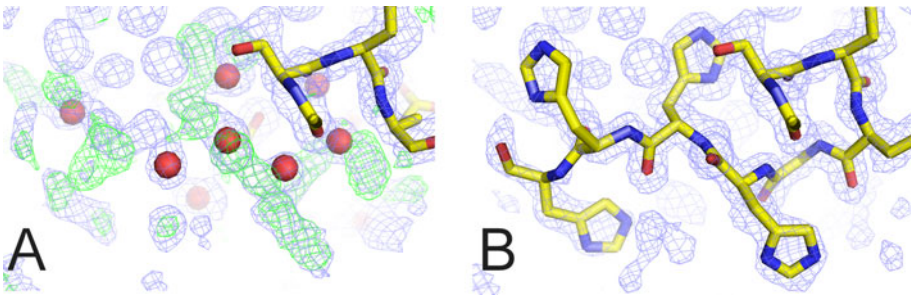

Fig. 6 Rerefinement of PDB structure 3BQS. **A** Electron density and model of 3BQS as reported in the PDB. **B** The same electron density region after re-refinement with HKL-3000 and redeposition as 3MAB

In general, automation has been critical to the success and high throughput of structural genomics. However, high throughput does not always translate into high output. As SG efforts increasingly focus on more difficult projects that require more flexible protocols, the automation pipeline requires expert intervention at critical decision points. The most successful approach is the development and automation of a multi-path approach that combines diversified protocols into an integrated and very efficient expert system.

Acknowledgments The authors would like to thank Zbyszek Dauter and Alex Wlodawer for valuable discussions; and Heping Zheng and Marcin Domagalski for help with generating statistics. The work described in the paper was supported by GM74942, GM53163 and with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200700058C. This work was supported in part by the U.S. Department of Energy, Office of Biological and Environmental Research and Office of Basic Energy Sciences, under contract DE-AC02-06CH11357.

References

1. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C (2002) The protein data bank. Acta Crystallogr D Biol Crystallogr 58:899–907
2. O'Toole N, Grabowski M, Otwinowski Z, Minor W, Cygler M (2004) The structural genomics experimental pipeline: insights from global target lists. Proteins 56:201–210
3. Wernimont A, Edwards A (2009) In situ proteolysis to generate crystals for structure determination: an update. PLoS One 4:e5094
4. Aslanidis C, de Jong PJ (1990) Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res 18:6069–6074
5. Dieckman L, Gu M, Stols L, Donnelly MI, Collart FR (2002) High throughput methods for gene cloning and expression. Protein Expr Purif 25:1–7
6. Camper DV, Viola RE (2009) Fully automated protein purification. Anal Biochem 393:176–181
7. Steen J, Uhlén M, Hober S, Ottosson J (2006) High-throughput protein purification using an automated set-up for high-yield affinity chromatography. Protein Expr Purif 46:173–178
8. Kim Y, Dementieva I, Zhou M, Wu R, Lezondra L, Quartey P, Joachimiak G, Korolev O, Li H, Joachimiak A (2004) Automation of protein purification for structural genomics. J Struct Funct Genomics 5:111–118
9. Viola R, Carman P, Walsh J, Frankel D, Rupp B (2007) Automated robotic harvesting of protein crystals-addressing a critical bottleneck or instrumentation overkill? J Struct Funct Genomics 8:145–152
10. Viola R, Carman P, Walsh J, Miller E, Benning M, Frankel D, McPherson A, Cudney B, Rupp B (2007) Operator-assisted harvesting of protein crystals using a universal micromanipulation robot. J Appl Crystallogr 40:539–545
11. Manjasetty BA, Turnbull AP, Panjikar S, Bussow K, Chance MR (2008) Automated technologies and novel techniques to accelerate protein crystallography for structural genomics. Proteomics 8:612–625

12. Zimmerman MD, Chruszcz M, Koclega KD, Otwinowski Z, Minor W (2005) The *Xtaldb* system for project salvaging in high-throughput crystallization. Acta Cryst A 61:c178–c179
13. Lesley SA, Wilson IA (2005) Protein production and crystallization at the joint center for structural genomics. J Struct Funct Genomics 6:71–79
14. Newman J, Egan D, Walter TS, Meged R, Berry I, Ben Jelloul M, Sussman JL, Stuart DI, Perrakis A (2005) Towards rationalization of crystallization screening for small- to medium-sized academic laboratories: the PACT/JCSG+strategy. Acta Crystallogr D Biol Crystallogr 61:1426–1431
15. Dong A, Xu X, Edwards AM, Chang C, Chruszcz M, Cuff M, Cymborowski M, Di Leo R, Egorova O, Evdokimova E, Filippova E, Gu J, Guthrie J, Ignatchenko A, Joachimiak A, Klostermann N, Kim Y, Korniyenko Y, Minor W, Que Q, Savchenko A, Skarina T, Tan K, Yakunin A, Yee A, Yim V, Zhang R, Zheng H, Akutsu M, Arrowsmith C, Avvakumov GV, Bochkarev A, Dahlgren LG, Dhe-Paganon S, Dimov S, Dombrovski L, Finerty P Jr, Flodin S, Flores A, Graslund S, Hammerstrom M, Herman MD, Hong BS, Hui R, Johansson I, Liu Y, Nilsson M, Nedyalkova L, Nordlund P, Nyman T, Min J, Ouyang H, Park HW, Qi C, Rabeh W, Shen L, Shen Y, Sukumard D, Tempel W, Tong Y, Tresagues L, Vedadi M, Walker JR, Weigelt J, Welin M, Wu H, Xiao T, Zeng H, Zhu H (2007) In situ proteolysis for protein crystallization and structure determination. Nat Methods 4:1019–1021
16. Kim Y, Quartey P, Li H, Volkart L, Hatzos C, Chang C, Nocek B, Cuff M, Osipiuk J, Tan K, Fan Y, Bigelow L, Maltseva N, Wu R, Borovilos M, Duggan E, Zhou M, Binkowski TA, Zhang RG, Joachimiak A (2008) Large-scale evaluation of protein reductive methylation for improving protein crystallization. Nat Methods 5:853–854
17. Walter TS, Meier C, Assenberg R, Au KF, Ren J, Verma A, Nettleship JE, Owens RJ, Stuart DI, Grimes JM (2006) Lysine methylation as a routine rescue strategy for protein crystallization. Structure 14:1617–1622
18. McPherson A, Cudney B (2006) Searching for silver bullets: an alternative strategy for crystallizing macromolecules. J Struct Biol 156:387–406
19. Derewenda ZS, Vekilov PG (2006) Entropy and surface engineering in protein crystallization. Acta Crystallogr D Biol Crystallogr 62:116–124
20. Chruszcz M, Wlodawer A, Minor W (2008) Determination of protein structures—a series of fortunate events. Biophys J 95:1–9
21. Soltis SM, Cohen AE, Deacon A, Eriksson T, Gonzalez A, McPhillips S, Chui H, Dunten P, Hollenbeck M, Mathews I, Miller M, Moorhead P, Phizackerley RP, Smith C, Song J, van dem Bedem H, Ellis P, Kuhn P, McPhillips T, Sauter N, Sharp K, Tsyba I, Wolf G (2008) New paradigm for macromolecular crystallography experiments at SSRL: automated crystal screening and remote data collection. Acta Crystallogr D Biol Crystallogr 64:1210–1221
22. Okazaki N, Hasegawa K, Ueno G, Murakami H, Kumasaka T, Yamamoto M (2008) Mail-in data collection at SPring-8 protein crystallography beamlines. J Synchrotron Radiat 15:288–291
23. Robinson H, Soares AS, Becker M, Sweet R, Heroux A (2006) Mail-in crystallography program at brookhaven national laboratory's national synchrotron light source. Acta Crystallogr D Biol Crystallogr 62:1336–1339
24. Muchmore SW, Olson J, Jones R, Pan J, Blum M, Greer J, Merrick SM, Magdalinos P, Nienaber VL (2000) Automated crystal mounting and data collection for protein crystallography. Structure 8:R243–R246
25. Cork C, O'Neill J, Taylor J, Earnest T (2006) Advanced beamline automation for biological crystallography experiments. Acta Crystallogr D Biol Crystallogr 62:852–858
26. Snell G, Cork C, Nordmeyer R, Cornell E, Meigs G, Yegian D, Jaklevic J, Jin J, Stevens RC, Earnest T (2004) Automated sample mounting and alignment system for biological crystallography at a synchrotron source. Structure 12:537–545
27. Cusack S, Belrhali H, Bram A, Burghammer M, Perrakis A, Riekel C (1998) Small is beautiful: protein micro-crystallography. Nat Struct Biol 5(Suppl):634–637
28. Pohl E, Ristau U, Gehrmann T, Jahn D, Robrahn B, Malthan D, Dobler H, Hermes C (2004) Automation of the EMBL hamburg protein crystallography beamline BW7B. J Synchrotron Radiat 11:372–377
29. Vernede X, Lavault B, Ohana J, Nurizzo D, Joly J, Jacquamet L, Felisaz F, Cipriani F, Bourgeois D (2006) UV laser-excited fluorescence as a tool for the visualization of protein crystals mounted in loops. Acta Crystallogr D Biol Crystallogr 62:253–261
30. Dauter Z (2005) Efficient use of synchrotron radiation for macromolecular diffraction data collection. Prog Biophys Mol Biol 89:153–172
31. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M (2006) HKL-3000: the integration of data reduction and structure solution–from diffraction images to an initial model in minutes. Acta Crystallogr D Biol Crystallogr 62:859–866
32. CCP4 (1994) The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr 50:760–763
33. Terwilliger TC, Berendzen J (1999) Automated MAD and MIR structure solution. Acta Crystallogr D Biol Crystallogr 55: 849–861
34. Terwilliger TC (2003) Automated main-chain model building by template matching and iterative fragment extension. Acta Crystallogr D Biol Crystallogr 59:38–44
35. Otwinowski Z (1991) ML-PHARE. in CCP4, SERC Daresbury Laboratory, Warrington, UK
36. Sheldrick G (2008) A short history of SHELX. Acta Crystallogr A 64:112–122
37. Perrakis A, Morris R, Lamzin VS (1999) Automated protein model building combined with iterative structure refinement. Nat Struct Biol 6:458–463
38. Cowtan K (1994) DM: an automated procedure for phase improvement by density modification. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallogr 31:34–38
39. Vagin A, Teplyakov A (1997) MOLREP: an automatic program for molecular replacement. J Appl Cryst 30:1022–1025
40. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53:240–255
41. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB III, Snoeyink J, Richardson JS, Richardson DC (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 35:W375–W383
42. Koclega KD, Chruszcz M, Zimmerman MD, Cymborowski M, Evdokimova E, Minor W (2007) Crystal structure of a transcriptional regulator TM1030 from Thermotoga maritima solved by an unusual MAD experiment. J Struct Biol 159:424–432
43. Long F, Vagin AA, Young P, Murshudov GN (2008) BALBES: a molecular-replacement pipeline. Acta Crystallogr D Biol Crystallogr 64:125–132
44. Keegan RM, Winn MD (2008) MrBUMP: an automated pipeline for molecular replacement. Acta Crystallogr D Biol Crystallogr 64:119–124
45. Lebedev AA, Vagin AA, Murshudov GN (2008) Model preparation in MOLREP and examples of model improvement using X-ray data. Acta Crystallogr D Biol Crystallogr 64:33–39
46. Schwarzenbacher R, Godzik A, Jaroszewski L (2008) The JCSG MR pipeline: optimized alignments, multiple models and parallel searches. Acta Crystallogr D Biol Crystallogr 64:133–140

47. Nienaber VL, Richardson PL, Klighofer V, Bouska JJ, Giranda VL, Greer J (2000) Discovering novel ligands for macromolecules using X-ray crystallographic screening. Nat Biotechnol 18:1105–1108
48. Terwilliger T (2004) SOLVE and RESOLVE: automated structure solution, density modification and model building. J Synchrotron Radiat 11:49–52
49. Terwilliger TC, Adams PD, Moriarty NW, Cohn JD (2007) Ligand identification using electron-density map correlations. Acta Crystallogr D Biol Crystallogr 63:101–107
50. Zheng H, Chruszcz M, Lasota P, Lebioda L, Minor W (2008) Data mining of metal ion environments present in protein structures. J Inorg Biochem 102:1765–1776
51. Allen FH, Taylor R (2004) Research applications of the cambridge structural database (CSD). Chem Soc Rev 33:463–475
52. Grabowski M, Chruszcz M, Zimmerman MD, Kirillova O, Minor W (2009) Benefits of structural genomics for drug discovery Research. Infect Disord Drug Targets 9:459–474
53. Read RJ, Kleywegt GJ (2009) Case-controlled structure validation. Acta Crystallogr D Biol Crystallogr 65:140–147
54. Urzhumtseva L, Afonine PV, Adams PD, Urzhumtsev A (2009) Crystallographic model quality at a glance. Acta Crystallogr D Biol Crystallogr 65:297–300
55. Myler PJ, Stacy R, Steward L, Staker B, Van Voorhis WC, Varani G, Buchko GW (2009) The seattle structural genomics center for infectious disease (SSGCID). Infect Disord Drug Targets 9:493–506
56. Anderson WF (2009) Structural genomics and drug discovery for infectious diseases. Infect Disord Drug Targets 9:507–517
57. Joosten RP, Womack T, Vriend G, Bricogne G (2009) Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. Acta Crystallogr D Biol Crystallogr 65:176–185
58. Koclega KD, Chruszcz M, Zimmerman MD, Bujacz G, Minor W (2010) 'Hot' macromolecular crystals. Cryst Growth Des 10: 580–586
59. Borrell B (2009) Fraud rocks protein community. Nature 462:970
60. Adams PD, Gopal K, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Pai RK, Read RJ, Romo TD, Sacchettini JC, Sauter NK, Storoni LC, Terwilliger TC (2004) Recent developments in the PHENIX software for automated crystallographic structure determination. J Synchrotron Radiat 11:53–55
61. Vonrhein C, Blanc E, Roversi P, Bricogne G (2007) Automated structure solution with autoSHARP. Methods Mol Biol 364: 215–230
62. Brunzelle JS, Shafaee P, Yang X, Weigand S, Ren Z, Anderson WF (2003) Automated crystallographic system for high-throughput protein structure determination. Acta Crystallogr D Biol Crystallogr 59:1138–1144
63. Holton J, Alber T (2004) Automated protein crystal structure determination using ELVES. Proc Natl Acad Sci USA 101:1537–1542
64. Panjikar S, Parthasarathy V, Lamzin VS, Weiss MS, Tucker PA (2005) Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. Acta Crystallogr D Biol Crystallogr 61:449–457