Published in final edited form as:

Expert Opin Drug Discov. 2011 August 1; 6(8): 771–782. doi:10.1517/17460441.2011.585154.

X-ray crystallography: Assessment and validation of protein-small molecule complexes for drug discovery

David R. Cooper^{1,2,4}, Przemyslaw J. Porebski^{1,3,4,5}, Maksymilian Chruszcz^{1,2,3}, and Wlodek Minor^{1,2,3,4}

¹ Department of Molecular Physiology and Biological Physics, University of Virginia, 1340 Jefferson Park Avenue, Charlottesville, VA 22908, USA ² Center for Structural Genomics of Infectious Diseases (CSGID), Jagiellonian University, Kraków 30-387, Poland ³ Midwest Center for Structural Genomics (MCSG), Jagiellonian University, Kraków 30-387, Poland ⁴ New York Structural Genomics Research Consortium (NYSGR), Jagiellonian University, Kraków 30-387, Poland ⁵ Department of Computational Biophysics and Bioinformatics, Jagiellonian University, Kraków 30-387, Poland

Abstract

Introduction—Crystallography is the key initial component for structure-based and fragment-based drug design and can often generate leads that can be developed into high potency drugs. Therefore, huge sums of money are committed based on the outcome of crystallography experiments and their interpretation.

Areas covered—This review discusses how to evaluate the correctness of an X-ray structure, focusing on the validation of small molecule-protein complexes. Various types of inaccuracies found within the PDB are identified and the ramifications of these errors are discussed. The reader will gain an understanding of the key parameters that need to be inspected before a structure can be used in drug discovery efforts, as well as an appreciation of the difficulties of correctly interpreting electron density for small molecules. The reader will also be introduced to methods for validating small molecules within the context of a macromolecular structure.

Expert opinion—One of the reasons that ligand identification and positioning, within a macromolecular crystal structure, is so difficult is that the quality of small molecules widely varies

Article Highlights

- The overall quality of the PDB is very good, but it should not be treated as an infallible repository. Poor quality structures pollute the PDB, and need to be filtered from structural data sets used to guide the drug discovery process. The derivation of principles that should govern drug design depends not only on the overall quality of structures within the PDB, but also the quality of the small molecules found therein.
- Structures must be validated before they can be used as a starting point for drug discovery efforts. Re-refinement may be necessary, especially for older structures.
- The drug discovery process is especially dependent on the correct identification and placement of small molecules within macromolecular structures.
- There is a dilemma of whether efforts should be put into the production of more data or into converting existing data into easy-to-use information.
- The conversion of information into knowledge would significantly increase the efficiency of structure-based drug discovery revolutionizing structure-based drug discovery.

Declaration of Interest The authors' research was supported 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 and also supported by NIH grants GM094662 and GM094585.

in the PDB. For this reason, the PDB can not always be considered a reliable repository of structural information pertaining to small molecules, and this makes the derivation of general principles that govern small molecule-protein interactions more difficult.

Keywords

Data mining; drug design; ligands; macromolecular crystallography; structural biology; structure validation

1. Introduction

Current biomedical science, including research performed in the pharmaceutical industry, is driven by high-throughput approaches. These approaches dramatically increase the amount of experimental data (for example, the rapid growth of macromolecular crystal structures as seen in Figure 1), but it is unclear if the increase in data has yielded a corresponding increase in useful information. The application of knowledge derived from structural data should yield improved computational methods that reduce the number of compounds to be tested experimentally, but most importantly should allow development of protocols to optimize the whole target-based drug discovery process. Unfortunately, the current ability of the drug discovery pipeline to transform structural data of drug targets into useful knowledge (and thus pay back the taxpayers' or stockholders' investments) does not meet the expectations of twenty years ago [1, 2]. This review focuses on the interpretation of structural data obtained by means of crystallography, rather than describing the details of the experimental methods need to produce, crystallize, and collect diffraction data on macromolecular crystals, which have been adequately reviewed in the recent literature [2-10]. One has to remember that an X-ray structure is merely a "snapshot" representing an ensemble average of a rigid macromolecule state, and sometimes it is useful to utilize other techniques better suited for measuring protein dynamics, such as NMR [7].

Modern drug discovery is based on a combination of several experimental and computational techniques. The ability of X-ray crystallography to provide accurate information about macromolecular structures, including molecular details of the interactions between proteins and small molecules, is unsurpassed. Consequently, crystallography is crucial for two lead generation paradigms within the pharmaceutical industry: it provides a starting point for computational structure-based drug design (SBDD), and is at the heart of fragment-based drug design (FBDD).

SBDD has long been used in the drug design industry [2, 11]. The first report of FBDD was published only a decade ago [12, 13], but it now plays a pivotal role in lead generation for most pharmaceutical companies [14]. Both SBDD and FBDD require diffraction-quality crystals of the drug target. This can be a formidable challenge in itself, but once the structure has been determined it needs to be accurately refined, because it will serve as a template for *in silico* screening (SBDD) or as a molecular replacement model (FBDD). This original structure does not need to be the *apo*-form of the protein; indeed, some proteins will not crystallize in the absence of co-factors or substrates. After refinement of the original structure, the SBDD approach can provide insight into the types of novel compounds that should be explored and reduce the number of compounds that need to be biochemically screened. However, obtaining the original structure is merely the first crystallographic step in FBDD, which involves soaking (or co-crystallizing) protein crystals in cocktails of small molecules that represent fragments of potential drugs. This, of course, adds the additional requirement of an abundant supply of protein and crystals with an accessible active site in the crystal lattice.

Typically FBDD libraries are relatively modest, with “only” several thousand compounds which are divided into cocktails containing a few compounds each. The compounds generally have a molecular weight between 100-250 Da and will have relatively low affinity for the protein, but can provide very detailed information on which chemical scaffolds do and do not specifically interact with the protein. This gives investigators a starting point for compound optimization [6]. Abbott Laboratories has reported a higher success rate for generating potent inhibitors using FBDD than using the more traditional high-throughput screening of chemical libraries, which requires significantly more screening experiments (by orders of magnitude) [14].

One of the crucial steps of FBDD is the initial interpretation of the electron density and modeling the chemical fragments into the active site of the protein. Theoretically, this task should be guided by rules about ligand-macromolecule interaction that can be determined by data mining of the PDB archive [15]. Overall, the quality of structures in the PDB is very good, but just as one bad apple can spoil a bushel, errors in structures can get propagated, especially when erroneous structures are used to derive rules that will influence subsequent model building. For example, even a simple query to determine the distribution of distances from metal ions to coordinating atoms in macromolecular structures returns ambiguous or even incorrect results [16]. In principle, these ambiguous results should not be related to the fact that the PDB contains structures determined at a wide range of experimental resolutions, as the chemistry of protein-ligand interactions is obviously resolution-independent. In practice, however, the resolution and crystallographic R-factors, which describe the agreement between the experimental diffraction data and the protein model, are used as the main filters to remove low quality structures. Indeed, in many cases applying resolution and R-factor filters are sufficient to obtain useful information on particular interactions, such as the distributions of metal-oxygen distances. One can easily verify these distributions by comparing them to those obtained from analysis of very high resolution X-ray structures of small molecules archived by the Cambridge Structural Database (CSD) [17]. Regrettably, these simple filters are not always sufficient to exclude all errors and thus produce a reliable reference set for protein/ligand interactions, and conclusions derived even from very high resolution protein structures can be completely misleading [18].

Occasionally the application of additional common sense rules is required to eliminate suboptimal structures from further analysis. Take for example the distribution of sodium ion to oxygen bond distances in PDB structures (Figure 2). Regardless of resolution, the distributions of bond distances deviate significantly from the distribution known from very high resolution small-molecule structures in the CSD. This disparity in bond distance distributions is even seen for the highest resolution PDB structures, which yields a distribution that is bimodal, contrary to both the CSD and known chemistry (Figure 2). However, either by excluding 2 PDB structures that contain implausibly high numbers of sodium atoms, or by excluding all PDB structures with >10 sodium atoms, one obtains distributions that correspond more closely to the analogous CSD distribution, regardless of resolution.

The development of general rules is even more complicated when one needs to estimate the quality of the placement and chemical correctness of ligands present in macromolecular structures. While X-ray diffraction is able to provide details of macromolecule-small molecule interactions, one should remember that the technique does not generate molecular models directly, but rather electron density maps that must be interpreted based on knowledge of protein-ligand interactions. An incorrect interpretation may significantly increase the amount of time between the identification of a drug target and the introduction of a drug to clinical trials.

2. Structure quality

High resolution diffraction data produce high quality electron density maps that allow relatively easy protein model building. However, the resolution and R-factors describe the overall quality of the structure, not the quality of local regions. Even high resolution maps can have regions that are difficult to interpret. These regions may correspond to the macromolecule itself or may be a small molecule that was introduced, intentionally or not, during protein production, crystallization and/or crystal harvesting. An unintentionally introduced small molecule may be an *in vitro* artifact or a ligand that is endogenously bound in the cell. This definitely complicates map interpretation, but may shed light on the location of an active site and/or other regions of the macromolecule with high propensity for ligand binding (Figure 3).

2.1 Validation tools for macromolecules

After a spate of incorrect structures was identified in the early 1990s [19], protein validation tools based on model geometry analysis were created that allowed the quality of the structure to be easily inspected and potentially erroneous residues identified. Programs such as PROCHECK [20] and WHATCHECK [19] provided crystallographers with an array of data that could indicate the overall quality of the structure, as well as an indication of the quality of the fit and geometry of individual atoms. The Ramachandran plot [21] is still considered one of the most informative checks of the overall quality of a protein structure because the dihedral angles of a protein measured by the plot are typically not restrained during refinement [22]. Protein structure validation was significantly advanced by MolProbity [23, 24], which supplements geometric and dihedral analysis with an all-atom contact analysis of hydrogen atoms within the protein structure, which has rapidly become an essential analysis. Indeed, the interactive molecular building program COOT [25] can now present MolProbity results and display the steric clashes instantly. An emphasis on protein model validation as a routine component of protein refinement, instead of just a last-minute check before structure deposition, has had a direct impact on structure quality since 2003 [23, 24]. In our opinion, failing to use macromolecular validation programs (or ignoring the serious warnings from them) before PDB deposition may pollute the PDB with inaccurate structures.

The current trend in structure validation is to combine the best parts of various validation programs under one umbrella. The new server PROSESS [26] claims to have the most extensive set of criteria to date for evaluating structures. It utilizes many well established programs and in addition can incorporate experimental NMR data. It also includes BLAST [27] searches and uses structural comparisons to identify potential problems. The PHENIX project also incorporates a validation tool that can evaluate a model using X-ray (and neutron) diffraction data. The graphical user interface of the PHENIX suite interacts with COOT (and MolProbity), so one can click on a problematic region identified by the validation tool and the COOT program will center on that region. The HKL-3000 suite [28, 29] also combines refinement and validation into a single process.

2.2 Analysis of structures

Once the model of a macromolecule has been refined and one is confident in its quality and chemical correctness, one may start analyzing the biological relevance of the structure. Is the biological unit seen within the crystal structure? This need not be within the asymmetric unit, since oligomerization can happen around a rotational symmetry element. Moreover, the structure should explain the protein's observed behavior and should be consistent with prior biochemical knowledge. The inconsistencies between a set of retracted ABC transporters and the wealth of biological data about them should have alarmed others [30].

Only after a model has been evaluated by all of the above criteria should it be analyzed for potential ligand binding sites. Once the binding sites are determined, properties like surface shape [31], electrostatics and atomic composition may be investigated and probed by computational or experimental approaches to identify lead compounds. However, even small errors either in the details of the original model or its interpretation may reduce the success rate of lead identification or slow down the process of structure-based drug design. It is difficult to overestimate the importance of structure quality, because a high quality X-ray structure of a macromolecule is only a starting point.

3. Validation of small molecules

Possibly the most crucial aspect of fragment-based drug design is the initial placement of the molecular fragments within the electron density. An incorrectly modeled chemical moiety at the onset of drug design process will at the very least hinder progress and can lead investigators down a dead end path. To further complicate matters, this initial placement can be difficult because non-covalently bound molecules can be more disordered than nearby residues, yielding hard-to-interpret electron density. Additionally, there may be more than one way to place the molecule within the electron density. Multiple examples of these types of difficult situations in fitting ligands to electron density were recently reviewed by Malde and Mark [32].

As of December 13, 2010, there were 60652 structures determined by X-ray crystallography in the PDB, and 83% percent of them contained at least one non-macromolecular and non-aqueous compound or atom. Overall there were 372643 small molecule chemical moieties (i.e. non-proteinaceous and non-aqueous moieties) in 50192 structures solved by X-ray crystallography. Validation tools for macromolecular structures are very well developed and widely used [19, 20, 33, 34]. Similarly the small molecule structures deposited to the CSD [17] are well validated, although the validation techniques used in small molecule crystallography are different due to the fact that all small molecule structures are solved to atomic resolution. Unfortunately only a very limited number of tools (e.g. PURY [35]) exist for validating small molecules or atoms in macromolecule-ligand complexes. For that reason, the quality of small molecules in protein complexes varies significantly from deposit to deposit [36, 37]. The validation must check the correlation between the electron density map and the small molecule model, and must also consider the geometry, atomic occupancy and B-factors. The validation also has to judge not only the overall chemical sanity of the small molecule, but also the chemical sanity of the environment. We would like to stress that an unusual chemical environment, which implies novel chemistry, has to be very well justified. It is very unlikely that medium or low resolution data (e.g. worse than 2.5 Å) alone will be sufficient to support the claim that novel chemistry was discovered.

3.1 Small molecule geometry

As stated above, the small molecule structures deposited to the CSD [17] are well validated and solved to atomic resolution. The lower resolutions, and subsequently the lower observation to parameter ratios, inherent in macromolecular structure refinements necessitate the use of geometric restraints derived from various sources. These restraints may be constructed from data mining of the CSD or calculated *ab initio* from theoretic values. Although there are several programs (like LIBCHECK [22], PRODRG [38], PURY [35], and PHENIX [39]) that automatically generate restraint sets for non-standard ligands, the restraints generated need to be carefully reviewed. Moreover, even very good restraints do not guarantee the proper geometry of the modeled small molecule. During refinement one can specify the relative weight of the model's geometric restraints versus the fit of the model to the experimental diffraction data in the maximum likelihood function used by the refinement algorithm. Very often this weight is controlled via a global parameter applied

equally to the protein and small molecule restraints, which is typically adjusted with respect to how the geometry of the protein deviates from ideality. However, a ligand which is less ordered than the macromolecule should use more heavily weighted geometric restraints, as compared to the restraints on the macromolecule, to maintain proper geometry. Ironically, the geometry of nonspecifically bound ligands in low-resolution structures is better than in medium-resolution structures because more heavily weighted restraints are applied during refinement [40].

3.2 Occupancy & B-factors

When inspecting a ligand that has been placed in the density, either manually or with the assistance of a program, attention should be paid to the atomic displacement parameters (commonly referred to as B-factors) and the crystallographic occupancy of the ligand or atoms. These two correlated crystallographic parameters are refined against experimental data and can provide insight into the quality of the small molecule within the context of the structure. However, currently there are no tools to (easily) check these parameters automatically, and thus the experimenter has to inspect them manually. Variable occupancy values within a molecule or significant differences in B-factors between neighboring atoms should raise suspicions about the correctness of ligand placement and refinement.

Since the B-factor of each atom is proportional to the uncertainty of its position, higher values indicate a higher degree of variation of the ligand's position, averaged over the different unit cells of the crystal and the duration of the diffraction experiment. The variation of B-factors within a ligand indicates how well-ordered each of the atoms are, with higher values associated with larger thermal vibrations and less precise positioning. Because of the way B-factors are calculated, they can grow quite large for incorrectly positioned atoms. The average B-factor of a noncovalently attached ligand should be reasonably close to, and certainly no more than double the value of, the B-factors of the atoms that contact the ligand.

3.3 Ligand – macromolecule interactions

As mentioned before, unless extremely high quality and high-resolution data is present, the probability of discovering novel or unusual geometry for a ligand is very low. Thus the ligands should closely resemble an idealized molecule, and any significant deviations should have an apparent explanation. Although very high resolution structures can occasionally allow for the determination of the protonation state of a molecule, typically the resolution limit of a macromolecular structure makes it impossible to directly detect hydrogen atoms. It is more common to have to resort to other indirect means to deduce the position of hydrogens. The most widely used method places hydrogens in their riding positions and optimizes the hydrogen bonding network to locate the rest. Hydrogen optimization is a crucial component of all-atom contact analysis, as calculated (for example) by MolProbity [24]. Information produced by MolProbity is very useful not only for validating macromolecules, but also provides insight into macromolecule-small molecule interactions. When provided prior information about the ligand, this suite can show hydrogen bonding patterns, steric hindrances and hydrophobic interactions. There are many other tools, especially those included in drug design suites and molecular modeling packages that assess and fix (to a limited degree) the ligand environment on the basis of potential energy. Unfortunately all of these calculations strongly depend on the pair wise interactions between the protein and ligand, and therefore a wrongly placed ligand may interfere with the proper refinement of the macromolecule.

3.4 Tools for ligand validation

Due to the increased importance of ligand identification and placement in structure-function relation studies and the drug discovery process, the number of tools that can assist crystallographers with ligand validation has increased. Recently, the PDB has started to validate ligands during deposition using simple constraints (e.g., bond lengths, angles, and chirality) and now provide depositors an opportunity to fix errors during the deposition process. To check a structure before deposition, the PURY [35] server will compare the bond lengths and angles to ideal values calculated from the CSD and identify atoms with improper chirality, flagging suspicious bonds, angles, and atoms. The ValLigURL server [41] provides similar information, and is linked with an experimental electron density server, allowing a direct comparison of a ligand in an experiment file to its corresponding electron density. The CheckMyMetal server can be used to validate metals within a macromolecular structure [3]. The PHENIX [42] validation tool mentioned above for macromolecules includes small model validation and can generate restraints for unknown ligands. However, similar to the restraints calculated by other software packages, these restraints need to be manually inspected.

Due to the difficulty of correctly placing small molecules within macromolecular structures, it was recently suggested [32] that molecular dynamics and energy minimization calculations should be routinely used in order to correctly determine the binding mode of small molecular components within macromolecular structures, even for structures determined to high resolution ($< 1.5 \text{ \AA}$). While there are certainly cases where the orientation of a small molecule within the electron density is ambiguous, several of the docking examples mentioned in their review would not have been as difficult if structural validation tools had been used to inspect the initial model of the protein for errors before attempting to dock a small molecule into poorly defined density. For example, when determining the tautomeric state of the ligand in the human cyclin dependent kinase 2 (CDK2) structure (1JVP), it could have been noticed that many of the residues were originally modeled with alternate conformations that were not justified by the electron density, but were necessary to avoid clashes with the alternate ligand conformation. After removing the alternate amino acid conformations and the complete ligand from the structure, re-refining and inspecting the residual density, it is clear that only one ligand conformation is present (Figure 4). Similarly, in the case of the cyclooxygenase-2 structure (1CX2), the overall geometry of the protein should have been better refined before attempting to address the assignment of nitrogen vs. oxygen in a sulphonamide group (Figure 4).

3.5 Missing or uninterpreted density

Currently there are three common practices for dealing with disordered fragments. The first practice is to model the small molecule in a “reasonable” conformation and let the B-factors rise and compensate for the disorder. The second practice is to assign the occupancy of the ill-fitting atoms to zero to denote that the atom cannot be reasonably modeled. It has to be stressed that atoms with occupancy values of zero are not refined against experimental data, and are therefore only a theoretical model. Some refinement programs will simply copy the coordinates of zero-occupancy atoms to the output file, even if their position makes no chemical sense or violates geometric restraints. Moreover, most molecular graphics programs display atoms with high B-factors or low occupancies just like any other atom. This can have the unintended effect of misleading non-crystallographers who view the structure but do not take the extra step of examining the occupancy. Therefore, this practice should be discouraged. In our opinion, the best practice is to set a firm standard for PDB deposits. Before standards are implemented, atoms that cannot be observed in electron density should be completely removed (which is the third practice). Removing atoms without sufficient electron density leaves the model incomplete, but it does not make any

assumptions about the conformation of the ligand or how it interacts with the protein, and clearly denotes the missing fragment as impossible to model on the basis of experimental data.

Some electron density maps contain uninterpreted regions. For example, we analyzed 43207 structures with electron density data provided by the Electron Density Server at Uppsala University [1] for unmodeled regions. Within the difference maps of 2915 structures, we found 33649 unmodeled density blobs too big to be water molecules ($> 50 \text{ \AA}^3$). An attempt to identify some of these unmodeled regions showed that in many cases the densities correspond well to small molecules or atoms found in the reported crystallization buffer. In some deposits even physiological ligands (e.g. uracil in 1PB6 and 3LOC) are omitted, despite the presence of clear electron density for them.

Fragments of unidentified densities can be clearly marked by modeling atoms (labeled “UNK” for unknown amino acid atoms or “UNL” for unknown ligand atoms) in those regions in the PDB deposition. Counterintuitively, the fraction of structures with unidentified atoms (corresponding to unattributed densities) is higher for higher resolution structures [43] because difficult-to-identify ligands are more easily detected and less likely to be incorrectly assigned in high resolution maps. Conversely, the interpretation of medium or low resolution maps is more subjective and makes it harder to validate the proper placement of small molecules. Sometimes unknown density regions are modeled as clusters of waters or not modeled at all. Unknown density may also be due to a poorly built or refined model or an error in space group assignment, which can make electron density map ligand interpretation very difficult [44]. When the experimental structure factors are deposited together with the protein model, the structure can be re-refined and reinterpreted, which may lead to correct identification of regions of unknown density (an example is PDB deposit 2NYD, a potential P_{II} regulator, which was redeposited as 3LNL after proper identification of unknown density as a molecule of the buffer bis-tris propane [45]) or transform a water cluster into a distinct protein fragment or ligand. It should be stressed that reinterpretation of a PDB deposit is almost pointless if the updated model is not redeposited to the PDB. In case of misidentification of the proper space group error [46], reprocessing of the diffraction data may be necessary, so the original data—in the form of raw diffraction images— should never be deleted or destroyed.

6. Conclusion

The growing number of structures in PDB should be a goldmine for large scale computational analysis. Unfortunately, structures in the PDB are not uniform in quality, as many structures are deposited by people with limited crystallographic experience. For example, 50% of all structures were deposited by people who are first authors of less than nine deposits. Moreover, the expectations of structure quality are much higher now than 15 years ago due to dramatic improvements in refinement programs and validation tools. Currently, large scale attempts to improve structures within the PDB are underway [47, 48], but automated protocols have their limitations, and some types of errors do not lend themselves to automatic correction.

Since the quality of refinement programs has significantly improved over older versions, similarly impressive improvements can be made to older structures simply by re-refining the structures (if the structure factors are available). Improved models may yield new structural details that are relevant from a biological point of view. In such cases, the re-refined models should be deposited to the PDB, and replace original deposits in order to provide the scientific community not only with better models, but also with a better starting dataset for bioinformatics analyses or model building. As we have indicated previously [18], the quality

of some structures used to train and develop docking programs is suboptimal, and these libraries could be significantly improved using the structure re-refinement/reinterpretation approach.

Structural information about the details of the molecular interactions derived from X-ray crystallography is very valuable, but this information is not sufficient to build a full picture of the process of molecular recognition between a protein and its ligand. Additional information derived from other experimental techniques like NMR, isothermal titration calorimetry (ITC), fluorescence- or spectrophotometry-based binding studies, mass spectrometry [49] or enzymatic assays is needed for a better understanding of the system. In addition to more routine experiments, neutron diffraction can provide crucial information on the protonation state of macromolecules [13, 41, 50].

7. Expert opinion

The problems encountered in structure-based drug discovery are similar to those in other fields of biomedical sciences and science in general. Despite the tremendous effort of many people, large monetary investments, and fantastic development of new technologies, the number of new drugs that has or will hit the market as the result of structure-based drug discovery is well below most expectations. The progress of molecular biology techniques [39] and the development of new crystallization technologies [51-54] has substantially shortened the time between target identification and the production of diffraction quality crystals. Substantial growth, particularly in the past five years, in the number of synchrotron stations dedicated for macromolecular diffraction experiments [55], the development of new detectors [56-59], and most of all, tremendous progress in software development [28, 42, 60] has brought an avalanche of new protein structures to the PDB. It is not uncommon for a structure to be solved before the experimenter finishes collecting the diffraction data, and structure refinement and deposition can take only a few weeks. For example, 11 percent of the structures solved by the four large-scale Protein Structure Initiative 2 (PSI2) centers were deposited within 14 days of completion of the diffraction experiment.

Unfortunately, the detailed analysis of the structure, *i.e.*, paper writing, takes much more time and (more often than not) is never completed. The difficulty of structure analysis can be illustrated by the fact that so far, only about 15% of Protein Structure Initiative structures are published in scientific journals, and a significant number of them are published in the form of “structure notes” which contain only a limited analysis. It is mandated that SG structures have to be deposited 6 weeks after refinement is complete, but most academic labs do not have that requirement. Many academic labs “sit” on refined structures until they are ready to publish the results, and as a consequence some structures fall through the cracks. For obvious reasons, countless structures solved by industrial labs never see the outside world.

With the exception of docking programs that use *ab initio* methods for restraint generation, the tools for *in silico* ligand docking [61] are only as good as the set of data used to develop them. The set of structures used to verify all of the program's assumptions must be carefully chosen. If that set contains even one structure with an incorrect ligand (*i.e.*, with insufficient electron density to justify its placement, its occupancy set to zero, etc.), in the best case unnecessary noise is introduced into the test set, and in the worst case, the flawed data may compromise the algorithms based on those sets. The errors proliferate since the docking programs are used by many researchers who blindly accept coordinates obtained from the PDB without evaluating the quality of the model. Therefore one should not be surprised that the reliability of *in silico* ligand screening is sometimes compared to the reliability of the

risk management programs that were used on Wall Street just before the 2008 stock market crash [43].

In general, errors in structures may proliferate very quickly. The structure of a β -glucosidase (PDB identifier 3FJ0) [62] which has 252 very likely misidentified sodium atoms has been downloaded from the PDB over 10000 times. Even taking into account that many of these downloads are performed by automatic procedures, this false data continues to widely proliferate. The sheer number of errors in public databases, which are scrutinized by the academic community, indicates that the proprietary databases used by commercial companies are also unlikely to be error free.

To provide the best possible source of data for the biomedical community, two new structural genomics centers—SSGCID [63] and CSGID [64], which concentrate on proteins from pathogenic organisms and complexes with small molecules—have established and published guidelines for the overall quality of the structures they deposit. In addition, the CSGID is using simple validation tools for structure quality and all deposited structures are divided into three categories: within the guidelines, significantly better than the guidelines, and significantly worse (<http://www.csgid.org/csgid/statistics/structures>). It is important to note that due either to the intrinsic flexibility of some parts of the protein or just because of poor resolution, there will always be structures of suboptimal quality. The structures that are significantly worse should not be used in global analysis, library generation or algorithm development, but they can still be very useful as starting points for pharmacophore identification, defining binding cavities, identifying important residues, and further experimental work.

What should be done with the 60000 X-ray structures that are already deposited? Most of them are very, very good, but even these structures can be slightly improved by using current software and our current knowledge. An example of this is the structure of lipoxxygenase; it was one of the highest quality structures in the PDB when it was deposited in 1996 [65]. Since then, we have re-refined and re-deposited the structural model three times (Table 1). The main goal of this exercise was to find what model improvement we can achieve using state of the art software and methodology. It is very interesting that even from a biomedical point of view, this was not an exercise for the sake of exercise - we corrected an error in the initial sequencing of the soybean lipoxxygenase [5]. The PDB_REDO project [66] does routine re-refinement of all PDB structures with both models and structure factors. However, this is a separate database that is unknown to many scientists that routinely use the PDB as a source of structural information. Moreover, because the re-refinement work is almost automatic, despite the improvement of basic global quality characteristics, more serious localized conceptual errors are still present in the re-refined structures. For example, in the re-refined deposit of 3FJ0 there are still 252 sodium ions, as in the original, and out of 79 deposits that reported zero occupancy for all ligand atoms, only 23 were corrected. This shows that automatic correction of models, as it is currently implemented, has limited applicability. Moreover, 10624 older deposits do not have structure factors and therefore cannot be corrected. This is the dilemma of current science: should we put effort into the production of more data or put some investment and effort into converting existing data into easy-to-use information? The conversion of information into knowledge is the most difficult step, but one that needs to be taken in order to significantly increase the efficiency of structure-based drug discovery. The success in this last step can revolutionize structure-based drug discovery and, together with non-crystallographic advances, can bring an avalanche of new drugs to market.

Acknowledgments

The authors would like to thank Matthew Zimmerman, Wayne Anderson and Alex Wlodawer for valuable comments on the manuscript.

References

1. Hol WG. Protein crystallography and drug design. *Arzneimittel-Forschung*. 1989; 39:1016–8. [PubMed: 2684178]
2. Congreve M, Murray CW, Blundell TL. Structural biology and drug discovery. *Drug Discovery Today*. 2005; 10:895–907. [PubMed: 15993809]
- 3•. Zheng, H.; Chordia, M.; Cooper, DR., et al. CheckMyMetal (CMM): Validation of metal binding sites in protein structures.. [In preparation. The announcement of a Web server for the validation of metals within macromolecular structures]
- 4•. Chruszcz, M.; Borek, D.; Domagalski, M., et al. Structural Genomics, Part C. Elsevier Academic Press Inc; San Diego: 2009. X-ray diffraction experiment - the last experiment in the structure elucidation process.; p. 23-40. [A description of experimental protocols used in collecting macromolecular X-ray diffraction data. This paper describes current experimental trends and provides statistical analysis of diffraction data reported to the PDB]
5. Chruszcz M, Wlodawer A, Minor W. Determination of protein structures - a series of fortunate events. *Biophysical Journal*. 2008; 95:1–9. [PubMed: 18441029]
- 6••. Hartshorn MJ, Murray CW, Cleasby A, et al. Fragment-based lead discovery using X-ray crystallography. *Journal of Medicinal Chemistry*. 2005; 48:403–13. [PubMed: 15658854] [An excellent review of the application of fragment-based drug design by Astex Pharmaceutical for several protein targets]
- 7•. Jhoti H, Cleasby A, Verdonk M, Williams G. Fragment-based screening using X-ray crystallography and NMR spectroscopy. *Current Opinion in Chemical Biology*. 2007; 11:485–93. [PubMed: 17851109] [Describes the benefits of combining X-ray crystallography and NMR to get a more comprehensive characterization of the active site]
8. Scapin G. Structural biology and drug discovery. *Current Pharmaceutical Design*. 2006; 12:2087–97. [PubMed: 16796557]
- 9••. Wlodawer A, Minor W, Dauter Z, Jaskolski M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *Febs Journal*. 2008; 275:1–21. [PubMed: 18034855] [A review discussing in detail all aspects of structure quality and the limitations of X-ray crystallography. A 'must read' for scientists who are entering the field of structural biology, or need to use structural biology results]
10. Carvalho AL, Trincão J, Romão MJ. X-ray crystallography in drug discovery. *Methods in Molecular Biology*. 2009; 572:31–56. [PubMed: 20694684]
11. Walkinshaw MD. Protein targets for structure-based drug design. *Med Res Rev*. 1992; 12:317–72. [PubMed: 1528078]
- 12•. Murray CW, Blundell TL. Structural biology in fragment-based drug design. *Current Opinion in Structural Biology*. 2010; 20:497–507. [PubMed: 20471246] [An excellent review of recent developments in fragment-based drug design with an emphasis on drug design for antimicrobial targets]
13. Nienaber VL, Richardson PL, Klighofer V, et al. Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nature Biotechnology*. 2000; 18:1105–8.
14. Hajduk PJ, Greer J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature Reviews*. 2007; 6:211–9.
15. Berman HM, Battistuz T, Bhat TN, et al. The Protein Data Bank. *Acta Crystallogr*. 2002; D58:899–907.
16. Zheng H, Chruszcz M, Lasota P, et al. Data mining of metal ion environments present in protein structures. *Journal of Inorganic Biochemistry*. 2008; 102:1765–76. [PubMed: 18614239]
17. Allen FH. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr*. 2002; B58:380–8.

18. Chruszcz M, Domagalski M, Osinski T, et al. Unmet challenges of structural genomics. *Current Opinion in Structural Biology*. 2010; 20:587–97. [PubMed: 20810277]
19. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. *Nature*. 1996; 381:272. [PubMed: 8692262]
20. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *J Appl Crystallogr*. 1993; 26:283–91.
21. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*. 1963; 7:95–9. [PubMed: 13990617]
22. Adams PD, Afonine PV, Bunkoczi G, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica*. 2010; 66:213–21.
23. Chen, VB.; Arendall, WB., 3rd; Headd, JJ., et al. MolProbity: All-atom structure validation for macromolecular crystallography.; *Acta Crystallographica*. 2010. p. 12-21. [New features of MolProbity are announced. MolProbity is available at <http://molprobity.biochem.duke.edu/>]
24. Davis IW, Leaver-Fay A, Chen VB, et al. MolProbity: All-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research*. 2007; 35:W375–83. [PubMed: 17452350]
25. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. *Acta Crystallogr*. 2010; D66:486–501.
26. Berjanskii, M.; Liang, Y.; Zhou, J., et al. PROSESS: a protein structure evaluation suite and server.; *Nucleic Acids Research*. 2010. p. W633-40. [The announcement of an excellent protein validation tool. The server is available at <http://www.prosess.ca>]
27. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *Journal of Molecular Biology*. 1990; 215:403–10. [PubMed: 2231712]
28. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M. HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. *Acta Crystallogr*. 2006; D62:859–66.
29. Cymborowski M, Klimecka M, Chruszcz M, et al. To automate or not to automate: this is the question. *J Struct Funct Genomics*. 2010; 11:211–21. [PubMed: 20526815]
30. Miller G. Scientific publishing. A scientist's nightmare: software problem leads to five retractions. *Science*. 2006; 314:1856–7. [PubMed: 17185570]
31. Nicholls A, McGaughey GB, Sheridan RP, et al. Molecular shape and medicinal chemistry: a perspective. *Journal of Medicinal Chemistry*. 2010; 53:3862–86. [PubMed: 20158188]
32. Malde AK, Mark AE. Challenges in the determination of the binding modes of nonstandard ligands in X-ray crystal complexes. *Journal of Computer-aided Molecular Design*. 2010 [Provides a discussion of the difficulties of the correct placement of small molecules within a macromolecular structures. Shows how molecular dynamics and free energy calculations can be used to determine the ligand's position when the electron density is ambiguous]
33. Kleywegt GJ. On vital aid: the why, what and how of validation. *Acta Crystallogr*. 2009; D65:134–9.
34. Kleywegt GJ, Henrick K, Dodson EJ, van Aalten DM. Pound-wise but penny-foolish: How well do micromolecules fare in macromolecular refinement? *Structure*. 2003; 11:1051–9. [PubMed: 12962624]
35. Andrejasic M, Praaenikar J, Turk D. PURY: a database of geometric restraints of hetero compounds for refinement in complexes with macromolecular structures. *Acta Crystallogr*. 2008; D64:1093–109.
36. Kleywegt GJ. Crystallographic refinement of ligand complexes. *Acta Crystallogr*. 2007; D63:94–100. [Required reading for practicing macromolecular crystallographers]
37. Evans PR. An introduction to stereochemical restraints. *Acta Crystallogr*. 2007; D63:58–61.
38. Schüttelkopf AW, van Aalten DM. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr*. 2004; D60:1355–63)
39. Moriarty NW, Grosse-Kunstleve RW, Adams PD. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta crystallographica*. 2009; 65:1074–80.
40. Gront, D. In preparation

- 41•. Kleywegt, GJ.; Harris, MR. ValLigURL: a server for ligand-structure comparison and validation.; *Acta crystallographica*. 2007. p. 935-8. [The totally cool ValLigURL server for ligand validation can be found at <http://eds.bmc.uu.se/eds/valligurl.php>]
42. Adams PD, Gopal K, Grosse-Kunstleve RW, et al. Recent developments in the PHENIX software for automated crystallographic structure determination. *Journal of Synchrotron Radiation*. 2004; 11:53–5. [PubMed: 14646133]
- 43••. Grabowski M, Chruszcz M, Zimmerman MD, et al. Benefits of structural genomics for drug discovery research. *Infectious Disorders Drug Targets*. 2009; 9:459–74. [PubMed: 19594422] [A review of the possible impact of SG programs on drug discovery research. This review provides insights into the strengths and weaknesses of SG programs and the role of such programs in drug development]
44. Choi J, Chon JK, Kim S, Shin W. Conformational flexibility in mammalian 15S-lipoxygenase: Reinterpretation of the crystallographic data. *Proteins*. 2008; 70:1023–32. [PubMed: 17847087]
45. Saikatendu KS, Zhang X, Kinch L, et al. Structure of a conserved hypothetical protein SA1388 from *S. aureus* reveals a capped hexameric toroid with two PII domain lids and a dinuclear metal center. *BMC Structural Biology*. 2006; 6:27. [PubMed: 17187687]
46. Poon BK, Grosse-Kunstleve RW, Zwart PH, Sauter NK. Detection and correction of underassigned rotational symmetry prior to structure deposition. *Acta crystallographica*. 2010; D66:503–13.
47. Joosten RP, Vriend G. PDB improvement starts with data deposition. *Science*. 2007; 317:195–6. [PubMed: 17626865]
48. Joosten RP, Womack T, Vriend G, Bricogne G. Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. *Acta Crystallogr*. 2009; D65:176–85.
49. Kim YJ, Doyle ML. Structural mass spectrometry in protein therapeutics discovery. *Analytical Chemistry*. 2010; 82:7083–9. [PubMed: 20698501]
50. Gibbs AC, Abad MC, Zhang X, et al. Electron density guided fragment-based lead discovery of ketohexokinase inhibitors. *Journal of Medicinal Chemistry*. 2010; 53:7979–91. [PubMed: 21033679]
51. Derewenda ZS, Vekilov PG. Entropy and surface engineering in protein crystallization. *Acta Crystallogr*. 2006; D62:116–24.
52. Dong A, Xu X, Edwards AM, et al. In situ proteolysis for protein crystallization and structure determination. *Nature Methods*. 2007; 4:1019–21. [PubMed: 17982461]
53. Koide S. Engineering of recombinant crystallization chaperones. *Current Opinion in Structural Biology*. 2009; 19:449–57. [PubMed: 19477632]
- 54••. McPherson A, Cudney B. Searching for silver bullets: an alternative strategy for crystallizing macromolecules. *Journal of Structural Biology*. 2006; 156:387–406. [PubMed: 17101277] [Discription of an alternative approach to the traditional sparse matrix or grid screening for initial crystallization conditions that includes mixing cocktails of small molecules with the protein]
55. Joachimiak A. High-throughput crystallography for structural genomics. *Current Opinion in Structural Biology*. 2009; 19:573–84. [PubMed: 19765976]
56. Broennimann C, Eikenberry EF, Henrich B, et al. The PILATUS 1M detector. *Journal of Synchrotron Radiation*. 2006; 13:120–30. [PubMed: 16495612]
57. Gruner SM, Tate MW, Eikenberry EF. Charge-coupled device area x-ray detectors. *Review of Scientific Instruments*. 2002; 73:2815–42.
58. Kraft P, Bergamaschi A, Broennimann C, et al. Performance of single-photon-counting PILATUS detector modules. *Journal of Synchrotron Radiation*. 2009; 16:368–75. [PubMed: 19395800]
59. Sakabe K, Sasaki K, Watanabe N, et al. Large-format imaging plate and Weissenberg camera for accurate protein crystallographic data collection using synchrotron radiation. *Journal of Synchrotron Radiation*. 1997; 4:136–46. [PubMed: 16699220]
60. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr*. 1994; D50:760–3.
61. Warren GL, Andrews CW, Capelli AM, et al. A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry*. 2006; 49:5912–31. [PubMed: 17004707]
62. Nam KH, Sung MW, Hwang KY. Structural insights into the substrate recognition properties of beta-glucosidase. *Biochemical and Biophysical Research Communications*. 2010; 391:1131–5. [PubMed: 20005197]

63. Myler PJ, Stacy R, Stewart L, et al. The Seattle Structural Genomics Center for Infectious Disease (SSGCID). *Infectious Disorders Drug Targets*. 2009; 9:493–506. [PubMed: 19594426]
64. Anderson WF. Structural genomics and drug discovery for infectious diseases. *Infectious Disorders Drug Targets*. 2009; 9:507–17. [PubMed: 19860716] [A description of the methods and approaches used by the Center for Structural Genomics of Infectious Diseases]
65. Minor W, Steczko J, Stec B, et al. Crystal structure of soybean lipoxygenase L-1 at 1.4 Å resolution. *Biochemistry*. 1996; 35:10687–701. [PubMed: 8718858]
66. Joosten RP, Salzemann J, Bloch V, et al. PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J Appl Crystallogr*. 2009; 42:376–84. [This project aims to provide an improved version of the PDB by re-refining structures in an automated way. While their approach certainly has its limitations, the comparison of the original PDB deposition with the PDB_REDO structure can provide an insight into the quality of the original structure]
67. Furet P, Meyer T, Strauss A, et al. Structure-based design and protein X-ray analysis of a protein kinase inhibitor. *Bioorganic & Medicinal Chemistry Letters*. 2002; 12:221–4. [PubMed: 11755359]
68. Kurumbail RG, Stevens AM, Gierse JK, et al. Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents. *Nature*. 1996; 384:644–8. [PubMed: 8967954]

Figure 1.

The number of new structures solved by X-ray crystallography deposited to the PDB per year. The yellow bars enumerate the total number of new X-ray deposits, the orange bars the number of new deposits that include experimental structure factors, and the blue bars the number of deposits that contain non-macromolecular (and non-aqueous) atoms and compounds.

Figure 2.

Distributions of oxygen-sodium bond distances for proteins in the PDB using different subsets of the PDB. Each row represents the distribution for a subset of the PDB partitioned by structure resolution limit (< 1.2 Å, 1.2-1.5 Å, 1.5-2.0 Å, and > 2.0 Å). The plots in column A represent all Na-O distances in all PDB structures that fall within each resolution range. The plots in column B use the same data as the plots in column A, except that they exclude two structures in the PDB (the β -glucosidase structures 3FJ0 and 3FIY) that contain implausibly many sodium ions (252 and 199 sodium ions, respectively). The plots in column C exclude all PDB structures within each resolution range containing greater than 10 sodium ions. For purposes of comparison, the Na-O bond distance distribution in the Cambridge Structure Database, which comprises atomic resolution X-ray crystal structures of small molecules, is shown as a black line.

Figure 3. Small molecules from the crystallization buffer may give insight into the active site. One of the first structures of dethiobiotin synthetase from *E. coli* (PDB code 1DBS, red) has sulfates (green) bound in the active site. The location of the bound sulfates was almost identical to the positions of the phosphate moieties in a cofactor molecule in one of the subsequent structures (1DAM, gray).

Figure 4.

A) A ribbon diagram of the human cyclin dependent kinase 2 (CDK2) structure (1JVP) is shown colored from blue (N-terminus) to red (C-terminus) [67]. Both conformations of the ligand PFK049-365 are shown and the positions of residues with alternate conformations are indicated by magenta spheres. B) After the removal of the ligand and all of the alternate conformations, the ligand was automatically added to the density using HKL-3000 and the structure was re-refined. For comparison, the previous positions of the ligand are shown (black). C) The Ramachandran plot of the cyclooxygenase-2 structure (1CX2) as calculated by MolProbity [30]. D) Structure 1CX2 [68] is shown as a ribbon diagram with Ramachandran outliers indicated by red spheres. All molecular figures were made using PYMOL [64].

Table 1

Comparison of refinement statistics for different models of soybean lipoxygenase-1. The models were obtained from the same dataset. The percentiles are calculated versus the same distribution of structures (from 2010).

PDB code	1YGE	1F8N	3PZW
Deposition year	1996	2000	2010
R (%)	19.7	20.5	14.9
R _{free} (%)	24.3	21.2	16.9
MolProbity score (percentile)	1.83 (51)	1.60 (74)	1.24 (96)
MolProbity clashscore (percentile)	9.12 (60)	7.56 (73)	3.76 (95)
Ramachandran favored (%)	95.0	96.9	97.7
Ramachandran additionally allowed (%)	4.3	2.7	2.3
Ramachandran outliers (%)	0.7	0.4	0
Rotamer outliers (%)	0.8	0.8	0.7