Chapter 21

The Quality and Validation of Structures from Structural Genomics

Marcin J. Domagalski, Heping Zheng, Matthew D. Zimmerman, Zbigniew Dauter, Alexander Wlodawer, and Wladek Minor

Abstract

Quality control of three-dimensional structures of macromolecules is a critical step to ensure the integrity of structural biology data, especially those produced by structural genomics centers. Whereas the Protein Data Bank (PDB) has proven to be a remarkable success overall, the inconsistent quality of structures reveals a lack of universal standards for structure/deposit validation. Here, we review the state-of-the-art methods used in macromolecular structure validation, focusing on validation of structures determined by X-ray crystallography. We describe some general protocols used in the rebuilding and re-refinement of problematic structural models. We also briefly discuss some frontier areas of structure validation, including refinement of protein–ligand complexes, automation of structure redetermination, and the use of NMR structures and computational models to solve X-ray crystal structures by molecular replacement.

Key words Structure quality, Structure validation, Drug discovery, Data mining, Structural genomics

1 Introduction

Structural genomics (SG) programs have greatly expanded our knowledge of the protein structure universe by determining almost 12,000 three-dimensional structures, which constitute approximately 14 % of the protein models that have been deposited to the Protein Data Bank (PDB) [1]. The NIGMS Protein Structure Initiative and NIAID Structural Genomics Centers for Infectious Diseases have alone supported determination of over 7,000 of these structures. However, a vast majority of them were not described in peer-reviewed articles and, taking into account the rate of new structures determined by SG, may never be published. Therefore, the scientific community will be able to access and evaluate them only through the data deposited in the PDB. For that reason the criteria that scientific community applies to model quality of SG structures should be stricter than for those coming from

Yu Wai Chen (ed.), *Structural Genomics: General Applications*, Methods in Molecular Biology, vol. 1091, DOI 10.1007/978-1-62703-691-7_21, © Springer Science+Business Media, LLC 2014

traditional structural biology laboratories. In addition, the overall quality of the deposits, including the completeness and accuracy of the header information, has to be as high as possible, since the remarks in the PDB files provide the only source of information describing the experimental methods that led to structure determination. Indeed, the average quality of 3D models coming from SG projects seems to be higher [2] than that of models coming from traditional structural biology laboratories. There are two reasons: (a) SG projects use very advanced technology and software tools, sometimes developed or enhanced by members of SG consortia; and (b) structural biologists at SG centers may be more experienced in structure determination, model refinement, and validation process than scientists working in traditional laboratories. Analysis of the authorship of PDB deposits shows that 54 % of all first authors served in this capacity for two or fewer deposits. So (perhaps) unlike peer-reviewed publication, PDB deposition seems to be an infrequent event in many biological laboratories. In this text, we discuss the impact of quality of structural models on biomedical research; in particular we address issues that are related to data mining and drug discovery research.

2 Protein Data Bank as a Data Mining Repository

2.1 Data Content of the PDB

The importance and role of the PDB for biomedical research cannot be overestimated. PDB is a unique repository containing atomic structural models of biological macromolecules (protein, DNA, and RNA) obtained by X-ray crystallography, NMR spectroscopy, electron microscopy, and other techniques. As 88 % of all PDB structures were determined by X-ray crystallography, our discussion of structural quality will focus mainly on this subset of the PDB. The PDB deposit for an X-ray diffraction structure usually contains three parts: (a) a header with information about diffraction experiment, structure determination, and refinement protocol; (b) coordinates of the atoms that make up the model of the macromolecules, water sites, and other small molecules in the structure; and (c) a structure factor file that contains diffraction data reduced from X-ray diffraction detector images.

In an ideal world, the first part (the header) would be equivalent to the "Materials and Methods" section in a typical peer-reviewed publication. However, the information in the headers of many PDB files is often contradictory, erroneous, and/or incomplete. The title of a PDB deposit is particularly important, especially if the deposit does not have a published citation, as it may be the only way to clearly identify whether the deposit is relevant to a given area of interest. Many structures have headers of the PDB files containing multiple values of "NULL," indicating that corresponding experimental data parameters are missing. The large

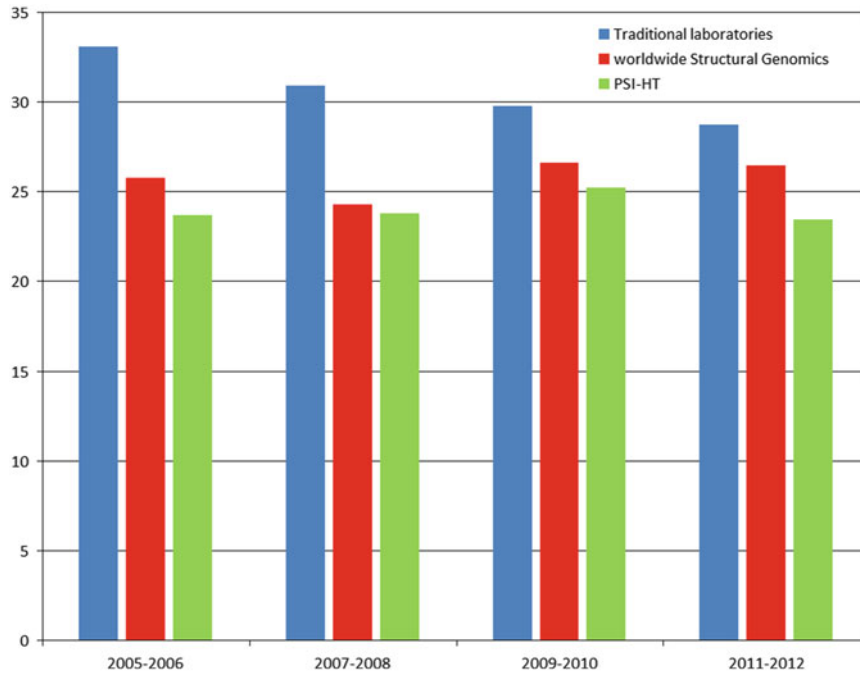

Fig. 1 Average number of missing parameters in the PDB file headers for PSI high-throughput (PSI-HT) centers, structural genomics worldwide (excluding the PSI-HT centers), and traditional structural biology laboratories. A small number of "NULL" values is always present due to generation of PDB file headers—not all parameters are relevant to all kinds of experiments

number of "NULL" data parameters should be alarming as it possibly indicates negligence and/or a lack of knowledge of how the crystallography experiment was performed. The number of "NULL" values for structural genomics centers is lower than average, not only because the depositors are more experienced, but because data needed for completing the header are usually readily available in, and possibly automatically extracted from, an existing database (Fig. 1).

The second part (coordinates of atoms in the macromolecular model) is usually the most reliable, as these coordinates are generated by refinement programs. However, there is no single standard for coordinate quality. For example, there are different methods for dealing with portions of crystallographically derived models corresponding to regions of weak or absent electron density. In some cases, all atoms of an amino acid residue are placed in probable locations regardless of density (with the occupancy parameters of atoms outside the map often reduced or set to 0). In others, atoms may be omitted from the model—perhaps only amino acid main chains are modeled, or only atoms unambiguously identifiable within the map are placed. Both approaches are justified for modeling uncertainty in experimental data, but can lead to very different results when thus derived coordinates are used as input to other

programs that calculate, for example, a charge on the surface of the protein. However, it should be noted that many, but not all, programs that read PDB files preprocess coordinates to address some of these ambiguities.

2.2 Inconsistent Quality of Models

Although the protein crystallography community (including structural genomics centers) has had many discussions about model quality, it has not agreed on a single, universal standard that models should meet before deposition. There are many quantitative measures that are clearly correlated with model quality, including resolution, R and R_{free} factors, distribution of deviations from ideal geometry, Ramachandran distribution, Molprobity clashscore, etc., but no single parameter is sufficient to conclusively determine whether a given structure is of high or low quality. "Quality" can also depend on context—the quality of a structural model useful for bioinformatics may be very different from its counterpart for in silico binding studies, for example.

As different depositors have different standards for deposition, mining of PDB data is very challenging. Fortunately, the PDB is unique among biomedical repositories as it contains experimental data as well. Since February 1, 2008 the PDB has required that each deposit based on crystallographic data must include a list of the structure factors used to build the model. When a structure is suspicious, in most cases a PDB user may download the corresponding structure factor file and re-refine the structure until it meets his or her own standards.

It is inevitable that there are differences in model quality standards since, to some degree the structures are based on subjective interpretation of experimental data. However, the X-ray diffraction models and experimental structure factor data in the PDB are generally of high quality, especially when compared to data in other repositories or databases used in biomedical research. In fact, the quality of the models and the ability of PDB users to examine and even re-refine a 3D model makes protein crystallography a "crown jewel" of experimental biomedical research.

Whereas there is some inconsistency in model quality due to a lack of universal deposition standards, much of this inconsistency is also due to the history of the field. For over 40 years more than 17,000 scientists have deposited models derived from experimental X-ray crystallography data of many different resolution limits, determined by various methods, and refined by many different, constantly evolving software packages. The distribution of high-resolution limits for all diffraction-based structures deposited in the PDB is very broad, as shown in Fig. 2. Even if the same software packages are used, quality of structures strongly depends on the design of diffraction experiments, data reduction, structure determination, refinement, and validation, particularly if multiple, weakly diffracting crystals are used. While the handling of diffraction

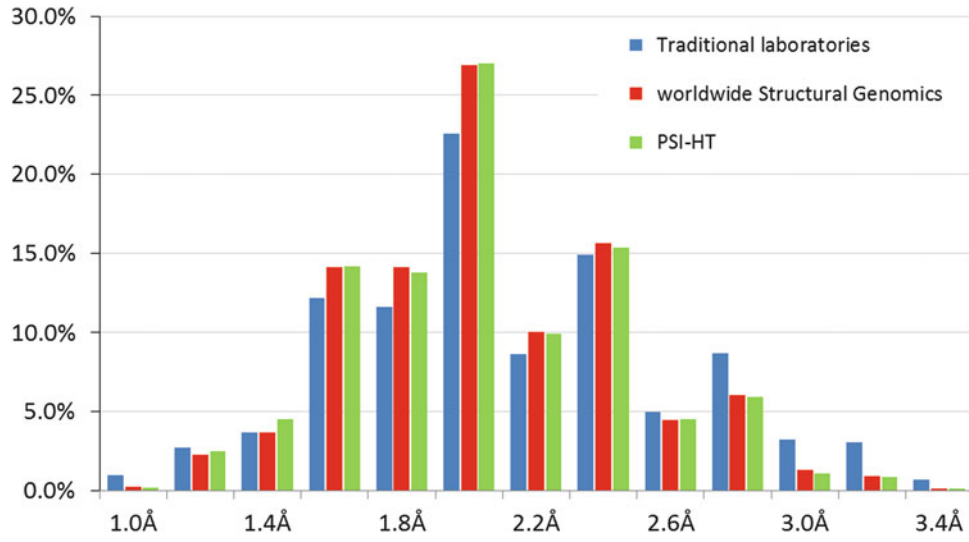

Fig. 2 Normalized distribution of high-resolution limits for X-ray structures solved by PSI high-throughput (PSI-HT) centers, structural genomics worldwide excluding PSI-HT, and traditional structural biology laboratories

experiments clearly depends on the experience and skills of the crystallographers performing them, examination of structures deposited by frequent depositors (i.e., those that are the first authors of more than 100 structures) shows that even different deposits prepared by the same person can vary significantly in quality measures (Fig. 3). Thus one has to acknowledge that the structure quality has to be also affected by factors other than experience, such as the quality of the experimental data. For example, models derived from poor resolution data—an intrinsic property of a crystal over which the crystallographer has little or no control—necessarily contain less information than a model from high-resolution data.

2.3 Selection of the Most Appropriate Deposits in PDB for Data Mining

All nontrivial data mining requires filtering and processing of the data in order to obtain reliable results. Much of this "quality control" work can be done in advance if there is curation, but most biomedical databases are either partially curated or not curated at all. PDB depositions are partially curated, as the authors receive extensive reports about problems in their depositions. The deposition reports produced by the PDB have steadily improved over the years, but there are still some areas for further improvement. For example, the current deposition tool (ADIT) does not yet validate the geometry of small molecules present in macromolecular crystal structures. Moreover, PDB depositors may ignore warnings in the report and ask that the model be deposited "as is." The most common protocol for filtering structures is defining a resolution limit cutoff for exclusion of lower resolution models from further analysis. In principle, this should be an ideal method, at least for the proteinaceous part of a macromolecular model. Unfortunately the

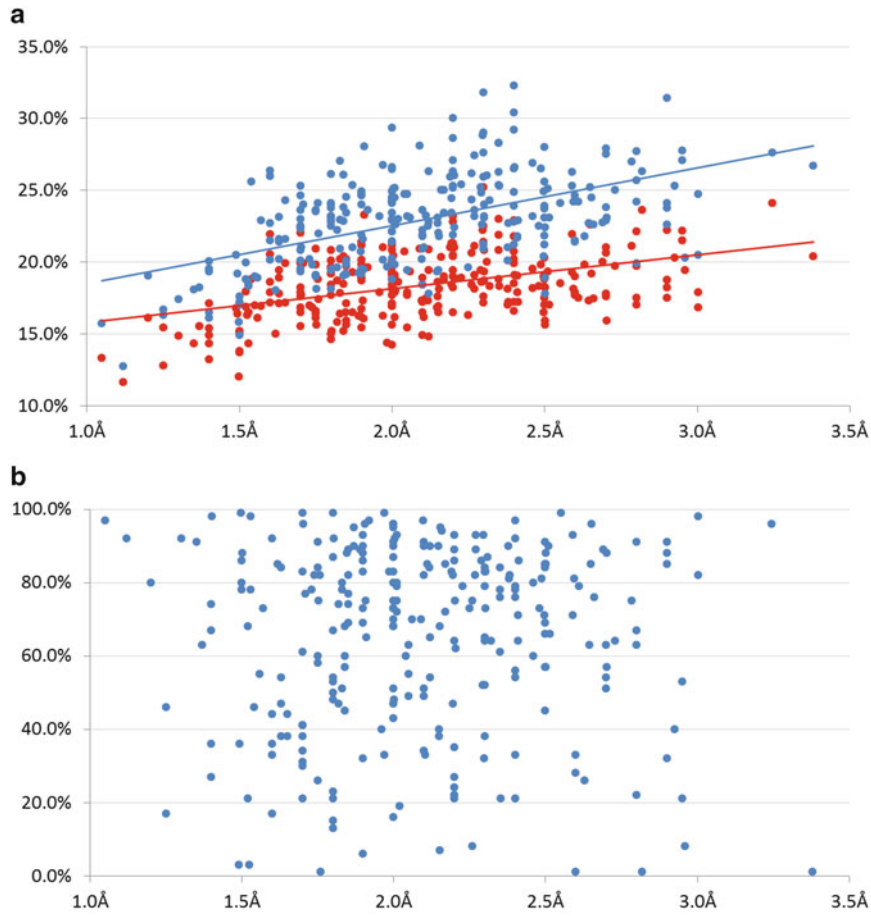

Fig. 3 Selected structure quality metrics of all PDB deposits with the same first author (the author was selected randomly from all such authors with >200 deposits). (**a**) Distribution of R (*red*) and R_{free} (*blue*) as a function of resolution, along with trendlines as determined by linear regression. (**b**) Distribution of Molprobity clashscore percentile (ranking of "raw" clashscore relative to other structures in the PDB of similar resolution)

high-resolution limit reported in a PDB deposit is not always equivalent to the nominal resolution limit of the diffraction data obtained from structure factors. In some cases, it appears that depositors may have chosen a resolution limit higher than is justified by the data. A significant number of PDB deposits include reflections in the highest resolution shell weaker, on the average, than the commonly accepted threshold (mean $I/\sigma(I) \geq 2.0$; *see* Fig. 4) [3]. (It should be noted that the traditional rule of "mean I over sigma ratio greater than 2.0" may not be the ideal way to choose a threshold; Karplus and Diederichs [4] have proposed an alternative statistic that advocates extension of the nominal resolution of a diffraction dataset.) However, analysis of the structure factor data in the PDB shows that the mean $I/\sigma(I)$ in the highest resolution shell of many diffraction datasets is as high as 10, suggesting that usable high-resolution reflections were never collected,

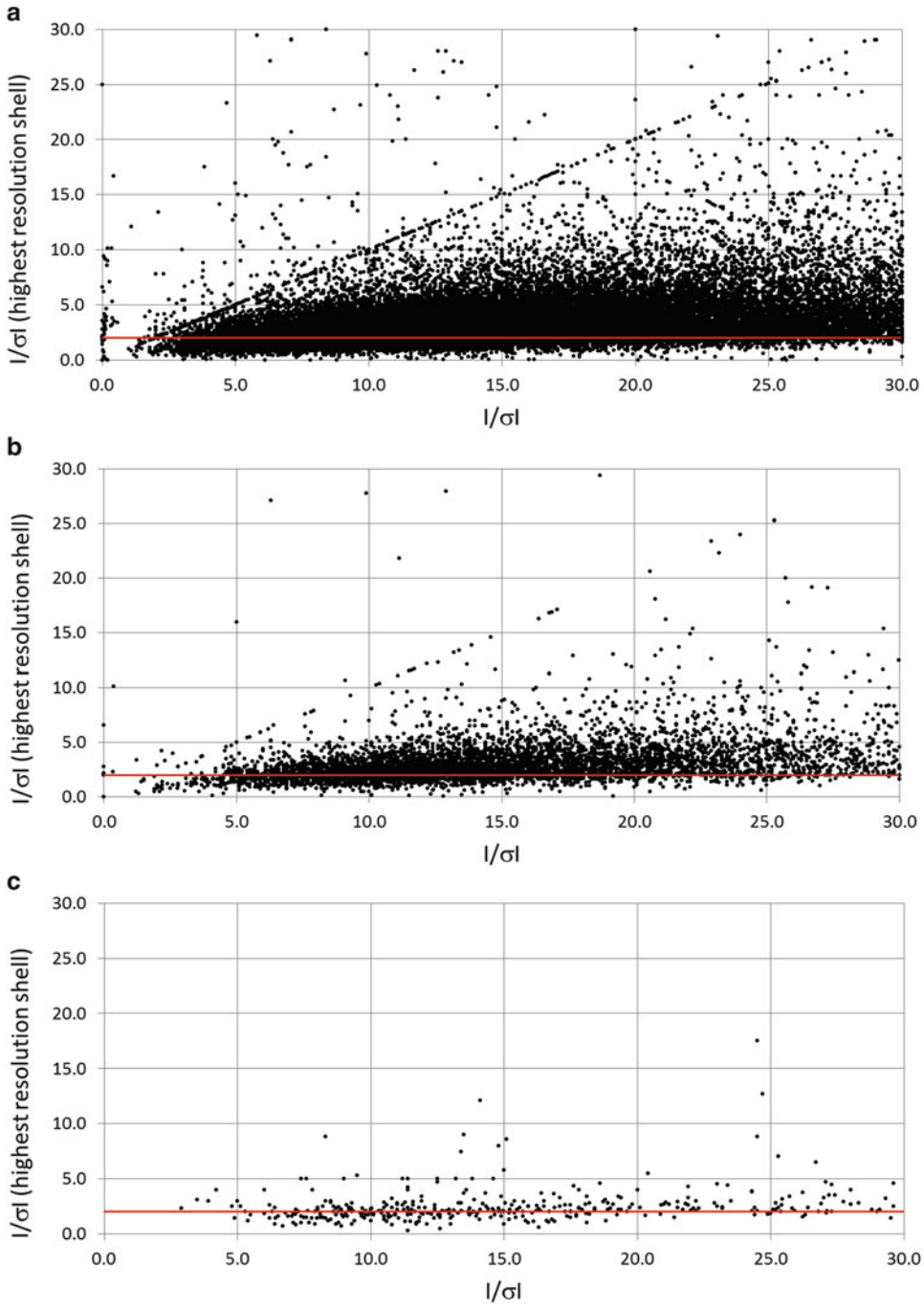

Fig. 4 Distributions of mean *I*/σ(*I*) for the highest resolution shell vs. mean *I*/σ(*I*) for all reflections, as determined for different sets of structures in the PDB. (**a**) Distribution for all structures determined by X-ray crystallography. (**b**) Distribution for all X-ray structures solved since April 2011. (**c**) Distribution for all X-ray structures solved since April 2011 by the four high-throughput PSI centers. On all distributions, the conventional threshold of 2.0 of mean *I*/σ(*I*) for the highest resolution is marked by a *red line*. There are a significant number of structures where the two values are identical, as well as a number where the mean *I*/σ(*I*) for the highest resolution shell is greater than the mean for all reflections, a physically improbable outcome

despite the tremendous investment of synchrotron beamlines in larger and faster detectors.

Moreover, dependent on the type of data mining analysis needed, validation may focus on either the macromolecular or small molecule portions of structures. In addition to evaluating the agreement between a structure model and the experimental data (R, R_{free}) and the properties of crystal packing (symmetry operations, solvent content), it is a common practice that structure determination, refinement, and validation in macromolecular crystallography are heavily (and necessarily) dependent on prior chemical knowledge of the subject molecule to define the geometry of the corresponding structural features in crystal structures [5]. There has been enormous progress in the development of statistics and tools used to verify the models of macromolecules in crystal structures, which measure agreement with both ideal geometry and experimental electron density. However, validation of small molecule models in macromolecular structures has lagged behind. Recently Rupp et al. demonstrated that a significant number of PDB deposits have ligands with very weak or even no correlation between the small molecule models and the electron density maps [6].

3 Model Quality

3.1 Overall Model Quality in the PDB

The validation tools developed over the years by many software authors [7–12], in addition to the in-house tool developed by the PDB [13], have greatly simplified the process of validation of protein models. Ideal values for bond lengths, bond angles, and dihedral angles within individual amino acid residues and in peptide bonds have been well defined and are incorporated into these programs [5]. Common secondary structural elements in protein structures (helices, strands, coils) can be defined by hydrogen bond patterns [14]. The overall geometrical quality of a protein main chain characterized by a Ramachandran plot [15] is particularly valuable for validation because the dihedral angles of individual peptides are usually not restrained during refinement. Potential steric clashes, which usually indicate problematic regions in a structure model, can be identified by Molprobity [12] and other similar programs. The agreement between diffraction data (structure factors) and a model are described by the R and R_{free} factors. PROSESS provides cross-validation with similar structures in the PDB to identify potential problems [16]. Despite the availability of a large selection of tools for structure validation, there is still no universal way to fully automate the process of model improvement. It is up to the crystallographer to utilize these tools routinely to identify potential problems and improve model quality after structure validation on a case-by-case basis, and it appears that nearly all follow this path. The

vast majority of models in the PDB are very good, despite the lack of precise definition of what values of the parameters describing structure quality are acceptable for high-quality structures.

Another often overlooked issue that may affect structure quality is that the structure factors are themselves derived quantities and thus do not represent the "raw" diffraction data used to determine a structure. Structure factors are typically reduced from a set of diffraction images collected in so-called rotation mode, and the way how the individual reflections on the images are indexed, integrated, and scaled together can significantly affect the quality of the structure factor amplitudes produced. Traditionally, the large size of diffraction image files has made it difficult to preserve (let alone distribute) raw diffraction data, but decreases in the cost per terabyte of hard drive storage have made storage and distribution of diffraction images feasible. Four SG centers, namely CSGID, SSGCID, MCSG, and JCSG, have made their diffraction images available for download from the respective servers. Diffraction images for over 2,200 PDB deposits, which comprise nearly 3 % of all X-ray structures in the PDB, are currently accessible. The public availability of original images provides an invaluable resource to determine if structure factors have been optimally reduced. The ability of the scientific community to access and evaluate raw, fundamental data directly from diffraction experiments makes crystallography arguably one of the most reproducible branches of biomedical science, with high transparency and reliability.

3.2 Quality of Macromolecular Structures Complexed with Small Molecules

Small molecules are abundantly represented in the PDB, as 80 % of PDB structures contain one or more residues that do not belong to polymers of amino acids or nucleic acids, or represent ordered water molecules. The presence of ordered small molecules in macromolecular structures usually highlights a specific area of interest or biological relevance. Although small molecules might be unintentionally introduced during sample preparation, the location of a small molecule in a macromolecular structure most often represents a binding site (or active site) that has some topological (concavity) or physiochemical properties suitable for binding. However, validation of small molecule models in protein structures is usually more difficult due to the diversity of small compounds and modes of interaction and conformation, i.e., the chemical sense of the environment. Moreover, even the use of high-resolution diffraction data does not necessarily guarantee high quality of the electron density around the small molecule, especially when there is always a possibility that the ligand may not fully occupy its binding site. For that matter, medium-to-low resolution of diffraction data is certainly insufficient by itself to justify the discovery of novel chemistry.

Small molecule models require specific tools to validate due to their chemical diversity and the fact that ligands are not covalently

bound to macromolecule, which can easily result in ambiguity in binding mode [17]. Geometrical parameters derived from the very high-resolution structures in the Cambridge Structure Database (CSD) [18] can be used as restraints in small molecule refinement [19], but in the case of a small molecule–macromolecule complex the procedures implemented to validate atomic resolution small molecule structures (such as the ones in the CSD) no longer apply. There are two main reasons for this. First, the overall resolution is usually significantly lower for a protein–small molecule complex compared to the crystal structure of a small molecule alone. Second, the binding mode of a small molecule needs to be validated, in addition to its conformation. Sometimes the models of small molecules are incomplete due to the degradation or multiple conformations. For that reason, the usage of stricter geometrical restraints is a common technique for the refinement and validation of small molecule binding sites in protein–small molecule complexes [20]. Therefore validation tools that can handle the small molecule portion of the complex are less common and often require substantial manual input to use. Consequently, the quality of small molecule models in PDB varies significantly. Useful tools for their validation include Twilight, which evaluates an agreement between small molecule models and electron density [6], and PURY, which evaluates the geometry [21].

4 Structure Validation

4.1 Validation Tools

Tools to validate structure quality, both overall and within substrate binding sites, are constantly evolving. However, the optimal ways of using these tools vary and are heavily dependent on the user's experience. For example, there is no common standard for a comprehensive set of parameters and threshold values to determine the validity of all structures. In addition, the standard protocols used for validation within most SG consortia are not yet streamlined or well defined. However, two SG centers, CSGID and SSGCID, have agreed that most of their targets should meet a common set of criteria. The structures determined within the HKL-3000 framework [22] may be checked by a standard validation procedure that compares quality parameters with the average values of the parameters as derived from structures deposited in the PDB during the last 2 years. Such a procedure was applied to and tested on more than 2,000 structures. The set of validation parameters, as implemented in HKL-3000 [22], could be easily applied to other software packages to standardize the validation process.

Structure validation is an ongoing, iterative process where model building and refinement are repeated until validation tools and visual inspection no longer reveal any problematic regions that can be further improved. However, no validation tools are perfect,

and none can objectively determine when a model cannot be further improved and should be considered "good enough." Therefore differences in knowledge and experience of crystallographers, or sometimes even just differences in opinion, may affect the decision whether or not a model is completed or should be refined further. The involvement of a second person to examine and evaluate the refinement of a structural model is usually considered a more objective approach for structure validation that can partially compensate the limits of experience and/or reduce the potential bias that a crystallographer may have during data interpretation. This approach is working successfully in a number of centers, including JCSG, NYSGRC, and the Structural Genomics Consortium (SGC).

As evaluated by R, R_{free}, Molprobity clash score, and Twilight score for ligands at a given resolution of structures, the average quality of structures determined by SG consortia in the PDB is significantly higher than the average quality of structures determined by other structural biologists over the last 2 years (Fig. 5). This trend is more prominent in overall quality assessment parameters such as the Molprobity clash score but is less prominent in ligand score, indicating that the availability of tools for a particular validation problem varies. Tools for overall validation, such as Molprobity or WHAT IF [23], have been available for a decade, whereas the tools for ligand refinement like Twilight were made available only recently. However, in the current year several SG structures of proteins complexed with small molecule ligands were redeposited, which suggests that SG efforts are promptly taking advantage of the new technologies.

4.2 Validation of Water Structure

As virtually all crystals of biological macromolecules are formed in aqueous solution, ordered water molecules bound to the surfaces of proteins and nucleic acids are commonly observed in X-ray crystal structures. However, at the resolutions of most macromolecular structures, typically only water oxygen atoms are observed. The binding of most waters is relatively weak. For example, NMR relaxation data show that nearly all protein surface water molecules have binding time scales of less than 100 ns, and molecular dynamics calculations predict residence times between 10 and 500 ps [24]. The residence time of even the most buried waters in a small protein BPTI was <20 ms [25]. The positions of most ordered crystallographic waters represent local energy minima into which waters fall reproducibly, appearing as peaks when averaged over all scattering events [26]. Only rarely are crystallographic waters in positions where they can form three or four H-bonds to other ordered atoms in the structure. It has also been noted that the number of crystallographic waters per residue identified in protein structures is inversely proportional to resolution [27, 28].

Although the binding of water molecules in the crystals is weak, their accurate modeling is still important for interpretation

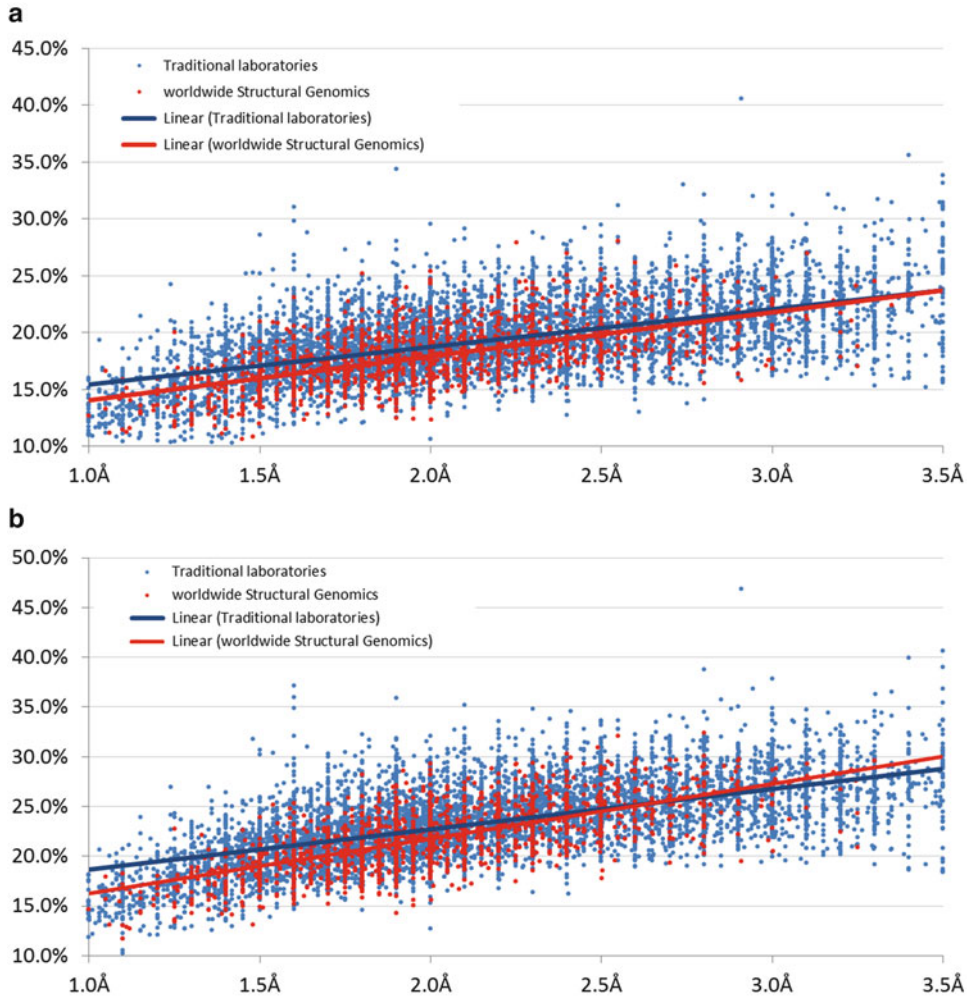

Fig. 5 (**a**) Distribution of *R* factor vs. resolution for all X-ray structures deposited in the PDB since April 2011. Structures solved by SG centers are marked in *red* and structures solved by traditional laboratories are in *blue*. The *lines* represent linear regression trend lines for the two sets of structures in the same color scheme. (**b**) Distribution of R_{free} factors vs. resolution for all X-ray structures deposited in the PDB since April 2011, using the same color scheme as part (**a**)

of results, because incorporation of ordered waters will improve the completeness of the model (and in turn, yield better estimates of the phases of the calculated structure factors). Crystallographically observed waters are not covalently bonded to the macromolecule in a structure, and at most resolution limits only a single peak corresponding to the oxygen atom is observed. Thus some spurious, "ghost" peaks in an electron density map can be mistakenly interpreted as waters, especially in medium-resolution structures. There are a number of tools for validating crystallographic water positions. For example, the interactive "Check Waters" tool in

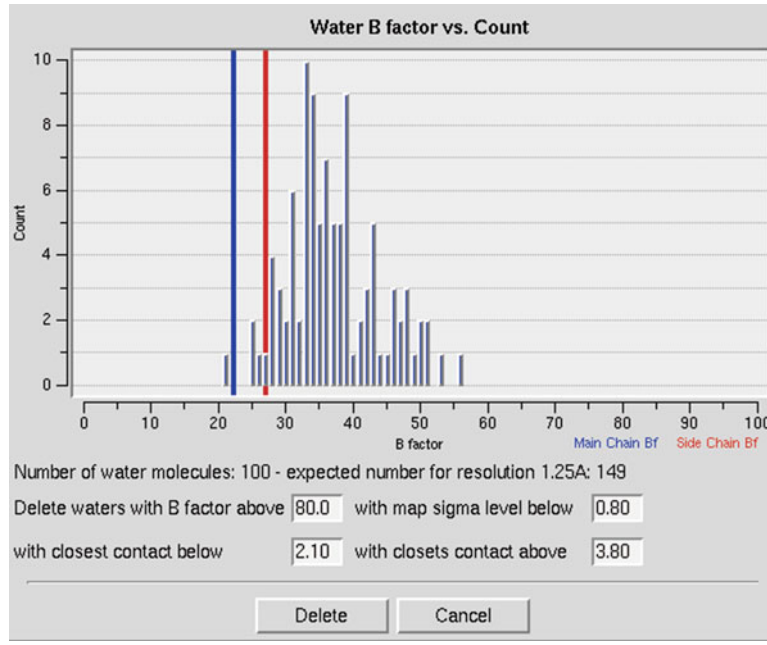

Fig. 6 A screen shot of the "Check waters" tool in HKL-3000

HKL-3000 [22] allows for effective validation of water molecules (Fig. 6). This is accomplished by plotting the distribution of waters as a function of atomic displacement parameters (or B-factors), providing information about the expected number of waters given the number of amino acids and resolution, following the method of Carugo and Bordo [27]. With this tool, all water molecules with B-factors greater than a user-defined threshold can be removed by one click.

5 Rebuilding and Re-refinement of Existing Models

5.1 The Benefits of Re-refinement

As mentioned above, an independent examination of a structure by a second researcher may reduce personal bias in data interpretation. In practice, availability of another expert to examine a structure is always limited, leading to the presence of a significant number of suboptimally refined structures in the PDB. For example, many structures that were determined in the past were refined and validated with tools that were quite primitive compared to the state-of-the-art tools in use today. Since many of these older structures describe important proteins and are frequently utilized as the basis for designing new experiments, it would be beneficial to revisit them using modern validation tools and reinterpret these structures more carefully. This would be especially instructive for structures that are used, for example, as test sets for in silico docking

experiments. Although, on average, SG-determined structures have relatively high structural quality, a routine re-refinement process is even more important because the consumer of a structure is likely to be less knowledgeable about X-ray crystallography and may take the structure "as is" in further biomedical research, e.g., as a target for structure-based drug design. Therefore deposition of SG structures of suboptimal quality will have a detrimental effect on subsequent research.

5.2 Automatic Re-refinement: PDB-REDO

One effort to re-refine old crystal structures with new technology on a large scale is the PDB-REDO project [29]. Each structure in the PDB for which structure factor data are available is automatically re-refined by a suite of tools using modern structure refinement and validation procedures, and, even more importantly, all of the different crystal structures processed by the system are handled uniformly, following a standardized refinement protocol. Even though the outcome of the refinement is still somewhat affected by the initial model, to a certain extent the PDB-REDO process removes the bias due to differences in refinement techniques used by different crystallographers. As a result, the quality statistics of the re-refined structures are more comparable. Advances in refinement techniques resulted in significant improvement of the refinement statistics, and in most cases, the values of R and R_{free} were improved by 2–5 %. However, PDB-REDO does not rebuild the original model (i.e., remove or add atoms other than in water molecules), which may be warranted if the electron density map is significantly improved. Whereas automated model building algorithms are becoming available, it has proven very difficult to fully automate this process with consistently reliable results [2]. Therefore, improvement in refinement protocols alone is not a panacea for maximizing the quality of crystal structures. As the majority of serious problems in structures that most affect structure quality require rebuilding of the model, large-scale automated re-refinement projects such as PDB-REDO are still limited. In addition, the PDB does not provide links to the PDB-REDO results. Researchers not familiar with structural biology are far more likely to use data from the PDB, so if PDB-REDO produces a model of higher quality from the same data, the biomedical community may never be aware of it.

In fact, inconsistencies between databases are some of the most significant impediments to effective biomedical data mining and research in general.

5.3 Semiautomatic Ligand Reassignment

As mentioned earlier, the potential presence of a small molecule constitutes a unique feature of the structure of an adjacent macromolecule. Given electron density of reasonable quality and the sequences of the polypeptides or nucleic acids, it is often relatively easy to build or rebuild the macromolecular portions of a structure, and in many cases this process can be automated (albeit with

human supervision) [30, 31]. However, correct identification and modeling of ligands is still difficult to automate, as the ligand bound is often unknown a priori and must be identified from an enormous and diverse set of endogenous substances. If the identity of the ligand is known (or limited to a small set), some tools such as RESOLVE [32], ARP/wARP [33], and the "Build ligand" tool in HKL-3000 [22] can search that set and automatically place a ligand in the map and refine it. Careful human examination of the search result is still crucial to verify correct placement. However, a search of a much larger chemical library is necessarily very computationally intensive and the approach described above does not scale. In contrast other programs such as PHENIX (phenix.ligand_identification) [34] or the MCSG-developed LigSearch [35], implement efficient protocols to search for potential physiological or drug-like small molecules in a much larger compound library. The potential ligands identified using a protein structure template may be very informative and may lead to the discovery of physiological ligands when unexplained electron density cannot be interpreted as compounds introduced during the processes of protein production or crystallization.

5.4 Structure Redetermination: Diffraction Images

Sometimes the structure factors that are deposited in the PDB are not sufficient to redetermine the structure. This is especially true when a structure was interpreted in the incorrect space group and the results affect the biomedical context of the structure. In such a case, the access to the original diffraction images is invaluable. Several years ago it was infeasible to store and distribute diffraction data, as building the storage and bandwidth infrastructure required to make diffraction data readily available to the research community was prohibitively expensive at best and impossible at worst. However, as storage media continue to rise in capacity and fall in price (as of this writing a 3 TB hard drive costs $130) and high bandwidth network connections are ubiquitous; the technical and financial barriers become less and less relevant. Application of efficient compression algorithms to diffraction images has further pushed the limits. However, the storage of thousands of datasets or more makes organization of data critical. As led by the four SG centers that make diffraction images available to the public (*see* Subheading 3.1 above), we may hope that, in the future, deposition of images in a public repository will become a requirement for publicly funded X-ray crystallography research. As shown recently, the possibility of reprocessing diffraction data that may not have been processed optimally (for example, by extending resolution limits and improving data quality) will lead to vastly improved models and their better interpretation [36]. Similar reprocessing will be beneficial in many ways, especially for relatively poorly determined structures.

6 NMR Structures

The use of NMR-derived models for solving crystal structures has been postulated and shown in practice over 20 years ago [37], but to date the success rate of such approaches has been somewhat limited. Many NMR models are simply not accurate enough to provide sufficient phasing power for the determination of crystal structures by molecular replacement. This is partly due to the nature of NMR data, which describe quite accurately local structures, but may not contain enough information to unambiguously assign long-range interactions. However, application of computational algorithms such as Rosetta has led to vast improvement in the success of molecular replacement calculations utilizing NMR models [38]. The Rosetta procedure is now part of standard crystallographic software [39]. As an example, it has been shown that its use, together with the involvement of computer games players [40], made it possible to utilize an NMR model for solving a structure by molecular replacement after many years of failure [41, 42].

7 Conclusions and Challenges

The current level of understanding of the biochemical mechanisms affecting living organisms would not be possible without the revolutionary progress of structural biology. The structures deposited in the PDB are only the starting point for many further analyses done by hundreds of thousands of scientists in academia and in industry. Any inaccuracy in a structure, even a small one, has tremendous potential to generate backlash, as the error may proliferate through all analyses that use data from that structure. In other words, a rotten apple can spoil the barrel. In addition, analyses of PDB structures can also be negatively affected by the lack of rigorous standards of data for PDB deposition. The X-ray diffraction structures determined by structural genomics centers worldwide are, on average, of higher quality than structures solved in traditional laboratories. Surprisingly, SG centers, who are unquestioned leaders in high-throughput and high-quality structure determination, have not established a precise definition of the conditions that could be universally used to assess the quality of macromolecular structures. Similarly, SG centers have not set a standard of deposition which could be adopted by the whole structural biology community. It is a serious challenge to establish both deposition standards and quality metrics, but large-scale SG projects are in a good position to propose them, due to the large databases that these efforts have generated. This is the key to success of all large-scale attempts to analyze the vast treasure which is the PDB.

Acknowledgments

The authors would like to thank Maks Chruszcz, Tom Terwilliger, Wayne Anderson, Matt Vetting, and Andrzej Joachimiak for valuable comments on the manuscript. This work was supported by PSI:Biology grants U54 GM094585 and U54 GM094662, as well as grant R01 GM053163, and supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research, as well as 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. HHSN272201200026C.

References

1. Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Res 28:235–242
2. Chruszcz M, Domagalski M, Osinski T et al (2010) Unmet challenges of structural genomics. Curr Opin Struct Biol 20:587–597
3. Grabowski M, Chruszcz M, Zimmerman MD et al (2009) Benefits of structural genomics for drug discovery research. Infect Disord Drug Targets 9:459–474
4. Karplus PA, Diederichs K (2012) Linking crystallographic model and data quality. Science 336:1030–1033
5. Engh RA, Huber R (1991) Accurate bond and angle parameters for X-ray protein-structure refinement. Acta Crystallogr A 47:392–400
6. Pozharski E, Weichenberger CX, Rupp B (2013) Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. Acta Crystallogr D 69:150–167
7. Laskowski RA, MacArthur MW, Moss DS et al (1993) PROCHECK—a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26:283–291
8. Hooft RW, Vriend G, Sander C et al (1996) Errors in protein structures. Nature 381:272
9. Oldfield TJ (1992) SQUID: a program for the analysis and display of data from crystallography and molecular dynamics. J Mol Graph 10:247–252
10. Pontius J, Richelle J, Wodak SJ (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 264:121–136
11. Vaguine AA, Richelle J, Wodak SJ (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr D 55:191–205
12. Chen VB, Arendall WB III, Headd JJ et al (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D 66:12–21
13. Yang HW, Guranovic V, Dutta S et al (2004) Automated and accurate deposition of structures solved by X-ray diffraction to the protein data bank. Acta Crystallogr D 60:1833–1839
14. Colloc'h N, Etchebest C, Thoreau E et al (1993) Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng 6:377–382
15. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7:95–99
16. Berjanskii M, Liang Y, Zhou J et al (2010) PROSESS: a protein structure evaluation suite and server. Nucleic Acids Res 38:W633–W640
17. Malde AK, Mark AE (2011) Challenges in the determination of the binding modes of non-standard ligands in X-ray crystal complexes. J Comput Aided Mol Des 25:1–12
18. Allen FH (2002) The cambridge structural database: a quarter of a million crystal structures and rising. Acta Crystallogr B 58:380–388
19. Lebedev AA, Young P, Isupov MN et al (2012) JLigand: a graphical tool for the CCP4 template-restraint library. Acta Crystallogr D 68:431–440
20. Gront D, Grabowski M, Zimmerman MD et al (2012) Assessing the accuracy of template-based structure prediction metaservers by comparison with structural genomics structures. J Struct Funct Genomics 13:213–225

21. Andrejasic M, Praaenikar J, Turk D (2008) PURY: a database of geometric restraints of hetero compounds for refinement in complexes with macromolecular structures. Acta Crystallogr D 64:1093–1109
22. Minor W, Cymborowski M, Otwinowski Z et al (2006) HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. Acta Crystallogr D 62:859–866
23. Vriend G (1990) WHAT IF: a molecular modeling and drug design program. J Mol Graph 8:52–56, 29
24. Brunne RM, Liepinsh E, Otting G et al (1993) Hydration of proteins. A comparison of experimental residence times of water molecules solvating the bovine pancreatic trypsin inhibitor with theoretical model calculations. J Mol Biol 231:1040–1048
25. Otting G, Liepinsh E, Wüthrich K (1991) Proton-exchange with internal water molecules in the protein BPTI in aqueous-solution. J Am Chem Soc 113:4363–4364
26. Bryant RG (1996) The dynamics of water–protein interactions. Annu Rev Biophys Biomol Struct 25:29–53
27. Carugo O, Bordo D (1999) How many water molecules can be detected by protein crystallography? Acta Crystallogr D 55: 479–483
28. Wlodawer A, Minor W, Dauter Z et al (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. FEBS J 275:1–21
29. Joosten RP, Joosten K, Murshudov GN et al (2012) PDB_REDO: constructive validation, more than just looking for errors. Acta Crystallogr D 68:484–496
30. Terwilliger T (2004) SOLVE and RESOLVE: automated structure solution, density modification, and model building. J Synchrotron Radiat 11:49–52
31. Perrakis A, Morris R, Lamzin VS (1999) Automated protein model building combined with iterative structure refinement. Nat Struct Biol 6:458–463
32. Terwilliger TC (2003) Improving macromolecular atomic models at moderate resolution by automated iterative model building, statistical density modification and refinement. Acta Crystallogr D 59:1174–1182
33. Langer GG, Evrard GX, Carolan CG et al (2012) Fragmentation-tree density representation for crystallographic modelling of bound ligands. J Mol Biol 419:211–222
34. Adams PD, Afonine PV, Bunkoczi G et al (2010) PHENIX: a comprehensive python-based system for macromolecular structure solution. Acta Crystallogr D 66:213–221
35. de Beer TA (2013) LigSearch. http://www.ebi.ac.uk/thornton-srv/databases/LigSearch/index.html. Accessed 23 Apr 2013
36. Perkins A, Gretes MC, Nelson KJ et al (2012) Mapping the active site helix-to-strand conversion of CxxxxC peroxiredoxin Q enzymes. Biochemistry 51:7638–7650
37. Baldwin ET, Weber IT, St Charles R et al (1991) Crystal structure of interleukin 8: symbiosis of NMR and crystallography. Proc Natl Acad Sci USA 88:502–506
38. Ramelot TA, Raman S, Kuzin AP et al (2009) Improving NMR protein structure quality by rosetta refinement: a molecular replacement study. Proteins 75:147–167
39. DiMaio F, Terwilliger TC, Read RJ et al (2011) Improved molecular replacement by density- and energy-guided protein structure optimization. Nature 473:540–543
40. Cooper S, Khatib F, Treuille A et al (2010) Predicting protein structures with a multiplayer online game. Nature 466:756–760
41. Khatib F, DiMaio F, Foldit Contenders Group et al (2011) Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nat Struct Mol Biol 18:1175–1177
42. Gilski M, Kazmierczyk M, Krzywda S et al (2011) High-resolution structure of a retroviral protease folded as a monomer. Acta Crystallogr D 67:907–914