The young person's guide to the PDB*

Wladek Minor[1,✉]

Zbigniew Dauter[2]

Mariusz Jaskolski[3,4]

[1]Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22908, USA
[2]Macromolecular Crystallography Laboratory, National Cancer Institute, Argonne National Laboratory, Argonne, IL 60439, USA
[3]Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University, Poznan, Poland
[4]Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

✉Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22908, USA; e-mail: wladek@iwonka.med.virginia.edu

*We dedicate this paper to Alex Wlodawer, an untiring ambassador of protein crystallography.

Received: February 7, 2016
Accepted: July 6, 2016

Key words: structural biology, macromolecular structure, Protein Data Bank, structural databases, structure validation, data mining

Acknowledgments: We would like to thank Heping Zheng and Ivan Shabalin for help and numerous discussions and suggestions. This work was supported in part by National Institutes of Health Grants HG008424, GM053163, GM117325, GM117080 as well as with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, under Contract No. HHSN272201200026C.

ABSTRACT

The Protein Data Bank (PDB), created in 1971 when merely seven protein crystal structures were known, today holds over 120,000 experimentally-determined three-dimensional models of macromolecules, including gigantic structures comprised of hundreds of thousands of atoms, such as ribosomes and viruses. Most of the deposits come from X-ray crystallography experiments, with important contributions also made by NMR spectroscopy and, recently, by the fast growing Cryo-Electron Microscopy. Although the determination of a macromolecular crystal structure is now facilitated by advanced experimental tools and by sophisticated software, it is still a highly complicated research process requiring specialized training, skill, experience and a bit of luck. Understanding the plethora of structural information provided by the PDB requires that its users (consumers) have at least a rudimentary initiation. This is the purpose of this educational overview.

INTRODUCTION

Structural biology is a young science, created ~60 years ago with the elucidation of the first macromolecular structures of the double helix of DNA in 1953 [1], and of the first proteins, myoglobin [2] and hemoglobin [3] at the end of 1950s. Parenthetically, we note that there is a big difference between these two sets of discoveries, not only because Watson and Crick used unacknowledged results of somebody else (X-ray photographs of Rosalind Franklin [4]) but mostly because the first DNA model provided just a general concept of this structure, while the two protein models consisted of accurate (within experimental error) coordinates of all the atoms (except hydrogen atoms) in three dimensions (3D). The atomic coordinates were derived from the information contained in the diffraction pattern recorded for the crystals of the macromolecule in question. Crystallography was therefore the first method that laid the foundation of structural biology, and it continues to be the mainstay of this information today. Initially, the crystallographic process was painfully long and tedious, and in the case of hemoglobin took Max Perutz 22 years of titanic work. Even in 1971, there were only seven protein structures known, all determined by X-ray crystallography [5]. That modest amount of biostructural information prompted, however, a visionary initiative to create a public repository of experimentally determined macromolecular 3D structures, under the name of the Protein Data Bank (PDB) [6]. The PDB holds today over 120,000 deposits, 90% of which come from crystallography. This incredible progress has been possible thanks to methodological advancements in physics and biology, dramatic increase of computer technology, and to progress of theory; still, however, the crystallographic process is far from an "automagic" one-button click and often requires a great deal of training, knowledge, skills, and sometimes a bit of luck from the practicing crystallographers. Training and adequate education are also needed for competent interpretation of the information stored in the PDB. Our article aims at providing a simple introduction to guide potential PDB users on their adventure into structural biology.

HOW A PROTEIN CRYSTAL STRUCTURE IS DETERMINED

In this article, we will use the time-sanctioned term "protein crystallography" with the understanding that it is an imprecise substitute for "macromolecular crystallography". A crystallographic experiment requires an X-ray source, a detector system that can measure the spatial distribution of diffracted X-rays, and a diffraction quality crystal. More than 99% of the deposited crystal structures were determined with the use of so-called single crystal rotation method [7]. The major obstacles for biologically relevant projects lie on the path: cloning-expression-purification-crystallization – from the genetic information to crystals. For well diffracting crystals, structure determination is usually straightforward and

can be accomplished using one of several semi-automatic packages, such as HKL-3000 [8] or PHENIX [9].

The diffraction experiments measure the intensities of diffracted X-rays, but cannot directly measure their phases, which are necessary to calculate electron density maps. The phases are primarily determined by one of two methods: (a) molecular replacement (MR) that uses starting phases from a similar model [10], and (b) multiple or single anomalous diffraction (MAD [11] or SAD [12]) that uses the anomalous signal of special atoms, such as selenium [13], to calculate experimental phases. A third method, multiple isomorphous replacement (MIR [14]), has fallen out of favor due to its reliance on toxic heavy metals and the technical difficulties involved. The isomorphous replacement method has been revived with the introduction of "quick halide soak" [15], which uses anions such bromide or iodide instead. In MR, it is very important to remove the model bias "inherited" from the model phases. In anomalous diffraction cases, it is possible to improve the experimental phases by solvent flattening [16] and the use of non-crystallographic symmetry (NCS) [17], which takes advantage of the fact that for many crystals there are more than one protein molecule in the asymmetric unit (ASU) of the crystal unit cell. The success of these methods is often beyond the experimenter's control, as the power of solvent-flattening depends on solvent content, and for crystals that have only one molecule in the ASU, the NCS map improvement cannot be applied.

It is important to realize that the final result of an X-ray diffraction experiment is an electron density map. Because of inadequate crystal quality (e.g., disorder or mosaicity), radiation decay, and imperfect experimental set-up, these maps may be noisy. Although the process of initial model building is usually performed by programs such as ARP/wARP [18], RESOLVE [19] or BUCCANEER [20], the final structural model is the crystallographer's interpretation of the electron density map. A noisy electron density map requires many iterations of manual intervention, and – in the extreme cases of very poor resolution – model building is entirely done 'by hand', using powerful molecular graphics software, such as COOT [21]. In the past (and unfortunately also at present in some crystallography laboratories) the model building step was followed by automatic model refinement, usually performed using REFMAC [22] or phenix.refine [23], and the validation process was delayed until the refinement was completed. Currently, all three steps, model building, refinement and validation (MBRV), are combined into a single-step process in sophisticated packages, such as PHENIX [9] or HKL-3000 [8]. However, one has to realize that even in the most productive and experienced crystallographic laboratories, an MBRV step can take between two hours and several weeks. Overall, the process involves scrupulous commitment to the iterative improvement and interpretation of the electron density maps.

One of the main difficulties of the refinement process (in which the model is optimized for its best consistency with the experimental diffraction data) is that protein crystals are not perfect and for that reason they usually do not diffract to very high resolution (Fig. 1). The consequence is not only noisy maps, but also the fact that the number of observables

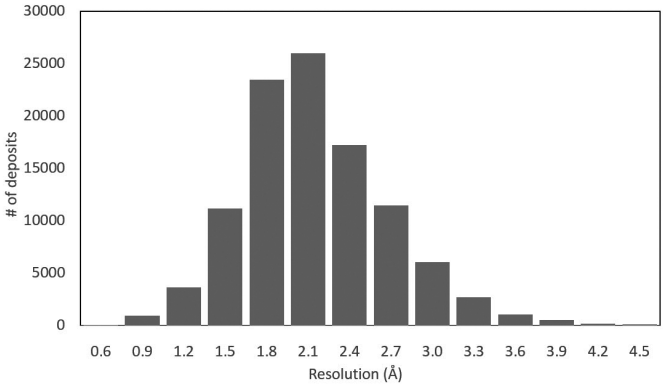

Figure 1. Histogram of the resolution of the structures deposited in the PDB during the last 5 years.

(i.e. diffraction peaks, or reflections) is not very high and frequently this number is comparable to or lower than the number of parameters necessary to correctly describe the model (the parameters are, for each atom: three geometrical coordinates (x,y,z) and the amplitude of vibration, often called the temperature factor B). For that reason, the refinement programs always use 'prior knowledge' about the geometry of the macromolecule, e.g. about typical bond distances and angles, that are obtained from high-resolution small molecule structures. The inability of computer programs (and crystallographers) to correctly interpret a noisy and/or weak electron density map is one of the main sources of errors in the models deposited in the PDB. An independent validation process that can be performed by various programs [24-26] and visual inspection of how well the model fits its corresponding electron density, may minimize the number and severity of errors, but do not eliminate them completely. It has to be accepted as a fact that any scientific model based on experimental evidence is also associated with a degree of error. The point is to be aware of this (consumers of the results) and make this degree as small as possible (creators of the results).

THE ANATOMY OF A PDB FILE

Nowadays, virtually all journals require depositions of structural results in the PDB prior to publication. Since 2007, submission of a structure to the PDB requires three components: (i) information about the crystal and experimental setup, (ii) the coordinates of all atoms (including ligands and solvent), and (iii) structure factors (i.e. intensity amplitudes) of all the diffraction spots measured during the experiment. The resulting "main PDB file" is produced in two forms. One is the so-called "Crystallographic Information File" or mmCIF (for example 2hyd.cif for the 2hyd deposit) that is mainly used for communication with other databases and computer programs. The equivalent simple text file (2hyd.pdb) has a header that contains information about the macromolecule and the authors, and about the experiment and structure determination, followed by a complete list of atomic coordinates and related parameters. In principle, the information that is provided in the header should be sufficient to write the 'Materials and methods' section of a paper describing the structure. However, in practice, the header

is often incomplete, erroneous, or even contradictory. For example, the header of the 2hyd deposit does not contain any information about how the correct structure was determined, as all fields that describe structure solution and the quality of the data are designated as 'NULL'. It is clear that nobody can correct the information in the header except the authors. The reason for inadequate headers is either negligence or lack of detailed knowledge about how the structure was determined. It seems that the latter occurs more often, as it has been demonstrated [27] that the degree of correctness of the header information is correlated with the correctness and accuracy of the structural model (atomic coordinates).

The coordinate section of the pdb file contains detailed information about each atom (protein, solvent or ligand) in the deposited model. Each atom is listed as a separate (ATOM) line that contains the following nine items: atom name, residue (amino acid) type, chain label, residue number, x, y, z coordinates (in Å), site occupancy, and ADP (atomic displacement parameter, in $Å^2$) also called the B-factor. In theory, the B-factor describes the magnitude of the atom's vibrations, but in practice it measures the accuracy of the atomic position. If the structure has been refined using anisotropic B-factors or TLS (Translation, Libration, Screw motion), each atom will have an additional line (ANISOU) with six parameters describing the anisotropy (different magnitude in three general directions) of the atomic vibrations. It is quite obvious that files that are many thousands of lines long are not analyzed by hand, but rather by one of advanced graphical programs, such as PYMOL (https://sourceforge.net/projects/pymol/) or COOT [21]. These and other programs display the structure, allow coordinate manipulation (geometry correction), and provide mechanisms to examine the three-dimensional model in a variety of ways, foremostly in confrontation with electron density maps, including aspects that are difficult to display, such as Ramachandran conformation [28], the analysis of B-factors, charge on protein surface, etc.

TRAPS AND OVERINTERPRETATIONS

The major difficulty of electron density map interpretation is associated with areas that are noisy and/or weak. The protein may have disordered parts, which will correspond to very weak and noisy parts in the electron density map. Model building in these areas is usually extremely difficult and even very experienced crystallographers may be only able to build "most probable" positions of the individual atoms. There is no consensus how to handle atoms that cannot be confidently modeled into electron density. One previously favored solution was to model those atoms into their most likely conformations, setting their occupancy to zero. Depending on the program used to display the model, this method can obscure the crystallographer's attempt to indicate disorder. Other crystallographers try to forcibly model and refine such atoms but this usually leads to poor geometry and "exploding" B-factors. Any atoms with B above ~60 $Å^2$ should be regarded as suspicious. Yet another approach to disordered fragments is to omit them altogether from the coordinate set. This will lead to incomplete models but in our opinion is the safest solution, clearly marking for the consumer of the model the problematic fragments. Besides, omitting "difficult" fragments is the first step of the calculation of so-called OMIT electron density maps (generated without contribution of the fragment in question), which very often will provide the best possible, unbiased view of the troubling area.

ERRORS: HOW TO DETECT AND AVOID THEM

The need for critical evaluation of and "limited trust" to the information contained in structural databases is highlighted by frankensteinase [29] (Fig. 2), a bogus enzyme with unrealistic features that are not likely to be found in any real protein structure. The problem with Fig. 2 (and with many crystal structure presentations) is that it is appealingly beautiful, and the aesthetic aspect tends to create the impression of correctness. Unfortunately, as in real life, good looks can be

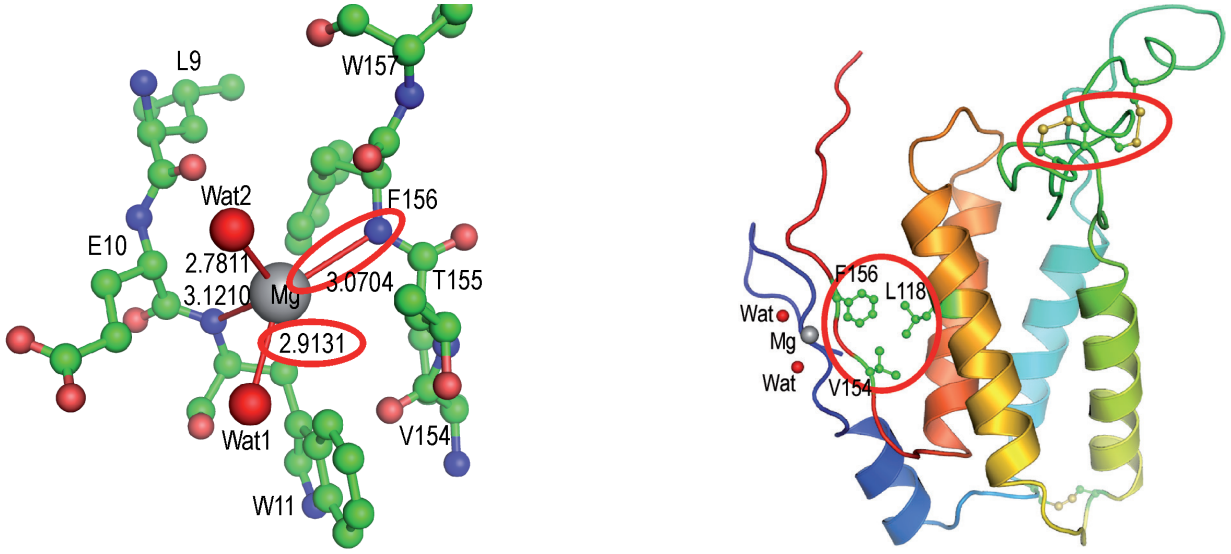

Figure 2. The structure of the enzyme frankensteinase in ribbon representation (right) and its purported metal "binding site" magnified on the left. The impossible features of the enzyme are highlighted by red ovals: (i) "active site" formed by non-polar residues; (ii) S-S bridges in a disordered region; (iii) metal "coordination" by amide N-H groups; (iv) exuberant precision (0.0001 Å) of unrealistic bond lengths (real Mg-O bonds are ~1.9 Å). The reader will be relieved to know that frankensteinase is a fake (albeit good-looking) "enzyme" constructed by Wlodawer *et al.* [29] for didactic purposes, by crude and creative stitching of bits and pieces from some real proteins taken from the PDB.

deceiving. One aspect that should alert the reader is that the bond lengths are presented with unrealistic precision that is quite impossible to achieve. This is one hallmark of overinterpretation: if the authors are unable to critically estimate the errors in their findings, the findings themselves are likely to be questionable as well. Fortunately, frankensteinase is only a "protein" fabricated for educational purposes, although it has a PDB file and all other attributes expected of a respectable scientific product. Unfortunately, there have been several instances of fraudulent data fabrications that made it to the PDB. While such cases could be potentially damaging to the reputation of the entire field of structural sciences, they had the beneficial effect of mobilizing the community for safeguarding its treasure [30] – the PDB – and also stimulated the creation of excellent validation tools [31] that can be used for constant lookout for potential errors in the PDB. As things stand now, it seems that the fabricated or forged models have been eradicated from the PDB.

(1) Apart from intentional forgeries, which hopefully are very rare, there are also other types of errors that occasionally may be found in the PDB. In the order of their seriousness they can be arranged in the following list:

(2) Totally wrong models, generated as honest errors. Such cases are very rare. One example is a series of ABC transporters [32,33], where the error has been later traced to local manipulation of the data processing program. When identified, the wrong models were retracted from the PDB.

(3) Incorrectly interpreted, otherwise decent data. An illustration is the structure of a RuBisCO subunit, which was traced backwards [34].

(4) Wrong connection between secondary structure elements.

(5) Register error, i.e. sequence shift during electron density interpretation.

(6) Wrong residue assignment. This error may be due to errors in sequence databases, to unexpected mutations, to mistakes during "electron-density sequencing", or to simple clerical errors (e.g., confusion between Asp/Asn).

(7) Wrong side chain conformation. This used to be a frequent error, but the model-building and structure validation tools (but not necessarily their users) are getting increasingly better in this respect. It is advisable to remember a simple stereochemical rule regarding the staggered (but not eclipsed) conformation around aliphatic single bonds: such torsion angles (e.g., C-C-C-C) can be ca. ±60° or 180°, but not ~120°.

(8) Misidentification of metal sites or confusing metal and water sites. The latter is especially frequent (and tricky) with isoelectronic species, such as $Mg^{2+}/Na^{+}/H_2O/NH_4^{+}$. As a guide, metal coordination bonds (e.g., Mg-O) are usually shorter than hydrogen bonds, are not formed by typical H-bond donors, such as the amide N-H group, and frequently are more numerous (e.g., six in octahedral coordination spheres) than the H-bond patterns formed by a water (or NH_4^{+}) molecule. It is good to remember that a water molecule is a double donor and double acceptor in H-bonding interactions, although bifurcated H-bonds (two acceptors interacting with an X-H donor) have to be taken into account.

(9) Unjustified solvent modeling. Adding too many water molecules and/or modeling them without support of electron density is a common sin of not only novices. As a rule of thumb, a structure may include $(3-|d_{min}|)$ water molecules per residue, depending on the resolution of the diffraction data (d_{min}, in Å). A water molecule should be included only if it is supported by good F_o-F_c electron density (at least 3σ), forms at least one decent hydrogen bond (O…O distance 2.3-3.2 Å), and does not have prohibitive (short) contacts with non-polar groups (although the existence of (typically quite long, C…O ~3 Å) C-H…O hydrogen bonds should not be overlooked).

(10) Unjustified or unreasonable modeling of atomic B-factors (ADPs) and/or occupancies. This can happen when individual anisotropic B-factors or overly complicated TLS models are refined at insufficient resolution, or when atomic occupancies (normally 1.0) are refined with disregard of logic and model parsimony (usually as a result of blind use of some programs).

(11) Unjustified modeling of H atoms. Hydrogen atoms are usually not located experimentally by protein X-ray crystallography (because they scatter X-rays very weakly) unless at subatomic resolution (0.9 Å or better). However, it is a good practice to include stereochemically generated H-atoms in structure factor calculations (F_c) even at medium resolution to improve agreement with F_o and to avoid an expanded skeleton syndrome. However, the H atoms should be riding on their parent atoms rather than being refined individually.

(12) A frequent error is fictitious modeling of features in electron density maps at too low contour level. It is often tempting to commit this seemingly trivial error of overinterpretation, especially when modeling ligand molecules, but the consequences could be very serious for the users of such models. For sound modeling in electron density, $2F_o$-F_c maps should be contoured at least at 1σ and F_o-F_c at 3σ.

The source of many of these errors is paucity of data (poor resolution) or their poor quality, very often resulting from suboptimal data collection/processing protocols. It can be difficult or impossible to collect more/better data once a diffraction experiment has been completed, which emphasizes the importance of checking for completeness and overall quality of a dataset during data collection.

We would recommend paying attention to the following warning signs of errors or problems in macromolecular models:

(1) High *R*-factor (>0.25).

(2) Large R_{free}-R discrepancy (>0.07). Large R_{free} [35] indicates inclusion of unjustified model parameters, i.e. overin-

terpretation. However, an R_{free} that is too close to R is also troubling, because it indicates that the test reflection have been compromised by inclusion (intentional or not) in model refinement. The test set must have reasonable size, usually ~1000 reflections. At low resolution it may be impossible to use R_{free} because there would not be enough reflections to sacrifice for testing. It is also not wise to waste too many (e.g., 5%) reflections when the data set is large (e.g., at very high resolution).

(3) Large deviations from standard stereochemistry. As an example, r.m.s. (root-mean-square) deviations from standard bond lengths above 0.03 Å should raise suspicion. However, very low (e.g., <0.001 Å) deviations may indicate that model restraints were weighted out of proportion at the expense of experimental observations. There is no point in imitating the stereochemical standards (restraints) better than justified by their own intrinsic inaccuracies. For the dictionary of standard bond lengths compiled by Engh and Huber [36,37] this would be ~0.02 Å. However, the newly proposed Conformation Dependent Stereochemical Library [38] may justify closer agreement.

(4) One should watch for signs of violation of chemical common sense. For example, a recent review has pointed out a number of chemical curiosities in the structures of protein complexes of cisplatin and carboplatin [39].

(5) Unusual B-factors, outside $4 < B < 60$ Å^2, or B-factors fluctuating wildly among connected atoms.

(6) Zero occupancy atoms, or fractional occupancies violating the covalent integrity of the model.

(7) Too many water molecules.

(8) Poor Ramachandran statistics (less than 90% in favored regions; disallowed angles).

(9) Unreasonable water/metal sites.

(10) Interpreting noise electron density. Sound modeling should withstand validation in $2F_o$-F_c maps at 1σ, or - better - in F_o-F_c OMIT maps contoured at the 3σ level.

(11) Reporting/discussing numerical values with unjustified precision, e.g. bond lengths to 0.0001 Å (vide supra).

Lack of experience leads many researchers to the conclusion that their structure is so special that it may violate many of the rules stated above. However, a discovery of such 'new chemistry' more often results in the replacement of the PDB deposit by the correct model, than in a phone call from Stockholm.

THE SPECIAL PROBLEM OF LIGAND MODELING

About 75% of the more than 120,000 PDB deposits include at least one ligand, and structural characterization of the interactions of macromolecules with small molecules can provide more information than a series of biochemical studies. It is thus obvious that macromolecular complexes with ligands are critical for drug discovery research [40]. However, analysis of the PDB shows that by almost any metric (geometrical and stereochemical quality, goodness-of-fit of the ligand to electron density, etc.), the quality of the ligands leaves a lot to be desired [41]. This is, no doubt, at least partly due to the fact that these ligands are not covalently bound to the protein, may have a higher degree of flexibility/disorder, and may not be present at full occupancy. Nevertheless, in many cases, a simple examination of the electron density map, or rather lack of density, shows that a 'ligand' molecule reflects only the wishful thinking of the researcher.

There are also many examples where continuous electron density has been filled with solvent molecules instead of, e.g. clearly identifiable polyhistidine tag (His_6) or a buffer molecule [42]. A His-tag may act as a competitive inhibitor of a peptide substrate and thus significantly affect the enzyme activity. Similarly, HEPES molecules from the buffer composition may bind in the substrate-binding site and influence the conformation of the active site. Thus, the history of the protein sample may significantly influence the course of its functional characterization. Reliable structural information could be crucial for proper understanding of such cases.

DATA MINING, THE EVOLUTION OF THE PDB AND OTHER RESOURCES

The PDB data can not only be used to visualize, examine, or analyze a single structure, but can also be used as a resource for data mining on a selected group of structural models. While detailed analysis of one particular structure is important for planning further biomedical experiments, data mining within the PDB can generate information that may impact the field as a whole. There are many papers that analyzed the structural models deposited in the PDB. However, even analysis of the metadata in the headers may also provide, for example, useful information for crystallization screen design and the way crystallization experiments are performed. An analysis of the available reports shows that until 2002, the sitting-drop experiments were used in less than 25% cases. After 2002, there has been a systematic increase of sitting drop experiments, and 2015 was the first year in which the number of sitting-drop experiments was higher than for hanging drop. The crystallographic liquid handling robots, such as Mosquito, and systems for automatic drop observation (e.g., RoboDesign or Formulatrix), are best suited for the sitting-drop setup. It is not clear whether the choice was dictated by convenience or careful consideration (or is of any consequence).

Initially, the PDB was designed as a resource for protein crystallographers, so the main page for each structure presented mostly information of interest to crystallographers, such as the crystallographic metrics and fold/motif descriptions. The growing impact of structural biology on biomedical sciences, including drug discovery, has influenced the transformation of the PDB into an extensive structural biology resource, and – during the last two years – into a thorough biomedical sciences resource.

For any data mining analyses, it is critical to have an error-free database. However, corrections of problems discovered in the PDB can take a significant amount of time - if they are made at all. Even if corrections are made to the PDB contents, disseminating the changes to other databases will take even more time. For that reason, the ripple effect of every error is faster than the ripple effect of its correction. The same can be observed with various scientific journals. Experience shows that the withdrawal of a paper with incorrect data can take a long time [43]. In the meantime, a lot of research effort can be ruined by overreliance on the incorrect data in the original publication. For that reason, we urge every user of any biomedical resource to view all information with a grain of salt. However, we are convinced that at present the PDB is the most reliable and most up-to-date among all biomedical resources.

OTHER METHODS CONTRIBUTING TO THE PDB

The vast majority of macromolecular deposits in the PDB come from single crystal X-ray diffraction experiments. The second largest group are structures from Nuclear Magnetic Resonance (NMR) experiments. NMR emerged in the 1980s as a powerful technique that can determine the solution structure of molecules smaller than ~400 kDa (today's limit). However, many researchers who have tried NMR models for Molecular Replacement found that despite the 'in principle' close homology, sometimes even 100% identity, the direct use of NMR models as MR probes did not lead to success. Those unsuccessful attempts led to the anecdotal deciphering of the NMR acronym as "Not for Molecular Replacement" [44]. Initially, the argument was that there might be a difference between macromolecular structures in crystalline form and in solution. However, the use of the Rosetta algorithm [45,46] to improve the NMR models proved to be very successful [47]. A spectacular example is provided by the structure of monomeric retroviral (M-PMV virus) protease. Its NMR model was unable to solve (by MR) the crystal structure. However, when the NMR model was corrected by the computer game players of Foldit [48] (which has a powerful Rosetta scoring algorithm) - the crystal structure could be finally determined [49,50].

Recently, cryo-EM (electron microscopy) has emerged as the most up-and-coming technique for determining the structure of large macromolecular complexes at resolutions comparable with X-ray techniques. Examination of the PDB shows an exponential growth of cryo-EM deposits [51] and one can expect even higher growth rates as the field matures. While cryo-EM is very promising (see the article by Czarnocki-Cieciura & Nowotny in this volume), it is not a silver bullet for structural biology; the difficulties of cryo-EM analysis were presented in a recent Cell article [52]. First, cryo-EM requires a sample with high molecular weight, usually above 200 kDa, although successful structure determination of proteins with sizes smaller than 100 kDa has been shown to be also possible, as a result of several innovations, such as the fixation of specimens with thin amorphous ice, ruthless selection of images, and the use of advanced direct electron detectors. Other limitations of cryo-EM include the relatively low resolution (which is defined in a somewhat fuzzy way) and the fact that the resolution of a given model is not uniform, i.e. worse at the outer regions that appear to be more flexible. Very few cryo-EM structures have resolution better than 2 Å. Despite of very fast progress, these limitations will make cryo-EM an unlikely technique in drug discovery research. For small proteins, improvements in resolution could be obtained by the use of Fab (antibody) fragments for complex formation [53]. The Fab fragments not only increase the size of the complex, but also assist with orientation assignment. The presence of one or two bound Fab fragments of ~50 kDa each, may bring almost any protein into the size range suitable for cryo-EM analysis. It appears that cryo-EM may converge in future synergistically with crystallography, where atomic models from X-ray crystallography would help to interpret the electrostatic potential maps (images) produced by cryo-EM. It has to be emphasized that electron microscopy is inherently related to electron diffraction, which is analogous to X-ray diffraction, except that the crystals can be (or even must be) very small, and the diffraction phenomena happen on the surface rather than in the bulk of the crystals (due to the very low penetration of electrons). Such electron diffraction studies of "invisible crystals" have been already reported [54]

CONCLUSIONS AND OUTLOOK

Each PDB consumer has to realize that there is no guarantee that a structure, even determined using high resolution data and with high-quality statistics (R_{free} and geometrical parameters) is error-free. Excellent R and R_{free} factors, as well as global model geometry, do not indicate that the local quality of the model is everywhere perfect. Structure refinement is a never-ending task, as improvements of the protocols and software for structure refinement/validation allow the creation of better models. Thus, a majority of the PDB models could be refined further, at least in theory [55], and such a practice may become the reality sooner than we think [56]. However, better structural models do not automatically mean better or more detailed biological information. For this to happen, one needs to use the brain a lot. But this is actually true of any step of scientific discovery.

This paper is dedicated to Alexander Wlodawer, a fantastic structural biologists and scholar but also a longtime personal friend. He can be compared to a new Hercules, who untiringly keeps cutting off Hydra's heads while new heads keep popping up. We do believe, however, that ultimately this won't be a Sisyphean task, and that his tenacious work and pressure will one-day lead to 'even more perfect' protein crystallography.

REFERENCES

1. Watson JD, Crick FH (1953) The structure of DNA. Cold Spring Harbor Symp Quant Biol 18: 123-131
2. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181: 662-666
3. Perutz MF (1960) Structure of hemoglobin. Brookhaven Symp Biol 13: 165-183
4. Franklin RE, Gosling RG (1953) Molecular configuration in sodium thymonucleate. Nature 171: 740-741

5. Protein Data Bank (1971) Protein Data Bank. Nature New Biol 233: 223
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242
7. Arndt UW, Wonacott AJ (1977) The Rotation Method in Crystallography. Amsterdam: North Holland
8. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M (2006) HKL-3000: the integration of data reduction and structure solution – from diffraction images to an initial model in minutes. Acta Crystallogr D62: 859-866
9. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D66: 213-221
10. Rossmann MG (1972) The Molecular Replacement Method. New York: Gordon & Breach
11. Hendrickson WA (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science 254: 51-58
12. Dauter Z, Dauter M, Dodson EJ (2002) Jolly SAD. Acta Crystallogr D58: 494-506
13. Hendrickson WA, Horton JR, LeMaster DM (1990) Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. EMBO J 9: 1665-1672
14. Green DW, Ingram VM, Parutz MF (1954) The structure of haemoglobin. IV. Sign determination by the isomorphous replacement method. Proc Roy Soc London A225: 287-307
15. Dauter M, Dauter Z (2007) Phase determination using halide ions. Methods Mol Biol 364: 149-158
16. Wang BC (1985) Resolution of phase ambiguity in macromolecular crystallography. Methods Enzymol 115: 90-112
17. Chapman MS (1998) Introduction to the use of non-crystallographic symmetry in phasing. In: Fortier S, ed, Direct methods for solving macromolecular structures. Kluwer, Dordrecht.
18. Perrakis A, Morris R, Lamzin VS (1999) Automated protein model building combined with iterative structure refinement. Nature Struct Biol 6: 458-463
19. Terwilliger TC (2003) SOLVE and RESOLVE: automated structure solution and density modification. Methods Enzymol 374: 22-37
20. Cowtan K (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr D62: 1002-1011
21. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr D60: 2126-2132
22. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D53: 240-255
23. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD (2012) Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr D68: 352-367
24. Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D66: 12-21
25. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A, Jones TA (2004) The Uppsala Electron-Density Server. Acta Crystallogr D60: 2240-2249.
26. Tickle IJ (2012) Statistical quality indicators for electron-density maps. Acta Crystallogr D68: 454-467
27. Grabowski M, Niedzialkowska E, Zimmerman MD, Minor W (2016) The impact of structural genomics: the first quindecennial. J Struct Funct Genomics 17: 1-16
28. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7: 95-99
29. Wlodawer A, Minor W, Dauter Z, Jaskolski M (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. FEBS J 275: 1-21
30. Minor W, Dauter Z, Helliwell JR, Jaskolski M, Wlodawer A (2016) Safeguarding structural data repositories against bad apples. Structure 24: 216-220
31. Read RJ, Adams PD, Arendall B 3rd, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Lutteke T, Otwinowski Z, Perrakis A, Richardson JS, Sheffler WH, Smith JL, Tickle IJ, Vriend G, Zwart PH (2011) A new generation of crystallographic validation tools for the Protein Data Bank. Structure 19: 1395-1412
32. Chang G, Roth CB, Reyes CL, Pornillos O, Chen YJ, Chen AP (2006) Retraction. Science 314: 1875
33. Matthews BW (2007) Five retracted structure reports: inverted or incorrect? Protein Sci 16: 1013-1016
34. Chapman MS, Smith WW, Suh SW, Cascio D, Howard A, Hamlin R, Eisenberg D (1986) Structural studies of Rubisco from tobacco. Phil Trans R Soc Lond B313: 367-378
35. Brunger AT (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 335: 472-475
36. Engh RA, Huber R (1991) Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr A47: 392-400
37. Engh RA, Huber R (2006) Structure quality and target parameters. In: Rossmann MG, Arnold E, eds, International Tables for Crystallography, volume F. Springer Netherlands, pp. 382-392
38. Tronrud DE, Karplus PA (2011) A conformation-dependent stereochemical library improves crystallographic refinement even at atomic resolution. Acta Crystallogr D67: 699-706
39. Shabalin I, Dauter Z, Jaskolski M, Minor W, Wlodawer A (2015) Crystallography and chemistry should always go together: a cautionary tale of protein complexes with cisplatin and carboplatin. Acta Crystallogr D71: 1965-1979
40. Zheng H, Hou J, Zimmerman MD, Wlodawer A, Minor W (2014) The future of crystallography in drug discovery. Expert Opin Drug Discovery 9: 125-137
41. Adams PD, Aertgeerts K, Bauer C, Bell JA, Berman HM, Bhat TN, Young J (2016) Outcome of the first wwPDB/CCDC/D3R Ligand Validation Workshop. Structure 24: 502-508
42. Majorek KA, Kuhn ML, Chruszcz M, Anderson WF, Minor W (2014) Double trouble-Buffer selection and His-tag presence may be responsible for nonreproducibility of biomedical experiments. Protein Sci 23: 1359-1368
43. Rupp B, Wlodawer A, Minor W, Helliwell JR, Jaskolski M (2016) Correcting the record of structural publications requires joint effort of the community and journal editors. FEBS J, in press, doi: 10.1111/febs.13765
44. Chen YW, Dodson EJ, Kleywegt GJ (2000) Does NMR mean "not for molecular replacement"? Using NMR-based search models to solve protein crystal structures. Structure 8: R213-220
45. DiMaio F, Echols N, Headd JJ, Terwilliger TC, Adams PD, Baker D (2013) Improved low-resolution crystallographic refinement with Phenix and Rosetta. Nature Methods 10: 1102-1104
46. Terwilliger TC, DiMaio F, Read RJ, Baker D, Bunkoczi G, Adams PD., Grosse-Kunstleve RW, Afonine PV, Echols N (2012) phenix.mr_rosetta: molecular replacement and model rebuilding with Phenix and Rosetta. J Struct Funct Genomics 13: 81-90
47. DiMaio F, Terwilliger TC, Read RJ, Wlodawer A, Oberdorfer G, Wagner U, Valkov E, Alon A, Fass D, Axelrod HL, Das D, Vorobiev SM, Iwai H, Pokkuluri PR, Baker D (2011) Improved molecular replacement by density- and energy-guided protein structure optimization. Nature 473: 540-543
48. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popovic Z (2010) Predicting protein structures with a multiplayer online game. Nature 466: 756-760

49. Khatib F, DiMaio F, Foldit Contenders Group, Foldit Void Crushers Group, Cooper S, Kazmierczyk M, Gilski M, Krzywda S, Zabranska H, Pichova I, Thompson J, Popovic Z, Jaskolski M, Baker D (2011) Crystal structure of monomeric retroviral protease solved by protein folding game players. Nature Struct Mol Biol 18: 1175-1177
50. Gilski M, Kazmierczyk M, Krzywda S, Zabranska H, Cooper S, Popovic Z, Khatib F, DiMaio F, Thompson J, Baker D, Pichova I, Jaskolski M (2011) High-resolution structure of a retroviral protease folded as a monomer. Acta Crystallogr D67: 907-914
51. Egelman EH (2016) The current revolution in cryo-EM. Biophys J 110: 1008-1012
52. Merk A, Bartesaghi A, Banerjee S, Falconieri V, Rao P, Davis MI, Pragani R, Boxer MB, Earl LA, Milne JL, Subramaniam, S (2016) Breaking cryo-EM resolution brarriers to facilitate drug discovery. Cell 165: 1698-1707
53. Wu YM, Chang JW, Wang CH, Lin YC, Wu PL, Huang SH., Chang CC, Hu X, Gnatt A, Chang WH (2012) Regulation of mammalian transcription by Gdown1 through a novel steric crosstalk revealed by cryo-EM. EMBO J 31: 3575-3587
54. Rodriguez JA, Ivanova MI, Sawaya MR, Cascio D, Reyes FE, Shi D, Sangwan S, Guenther EL, Johnson LM, Zhang M, Jiang L, Arbing MA, Nannenga BL, Hattne J, Whitelegge J, Brewster AS, Messerschmidt M, Boutet S, Sauter NK, Gonen T, Eisenberg DS (2015) Structure of the toxic core of α-synuclein from invisible crystals. Nature 525: 486-490
55. Joosten RP, Joosten K, Cohen SX, Vriend G, Perrakis A (2011) Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. Bioinformatics 27: 3392-3398
56. Minor W, Dauter Z, Helliwell JR, Jaskolski M, Wlodawer A (2016) Safeguarding structural data against bad apples. Structure 24: 216-220

Wprowadzenie młodego człowieka do PDB*

Wladek Minor[1,✉], Zbigniew Dauter[2], Mariusz Jaskolski[3,4]

[1]University of Virginia, Charlottesville, VA 22908, USA
[2]Macromolecular Crystallography Laboratory, National Cancer Institute, Argonne National Laboratory, Argonne, IL 60439, USA
[3]Zakład Krystalografii, Wydział Chemii, Uniwersytet im. A. Mickiewicza w Poznaniu ul. Grunwaldzka 6, 60-780 Poznań, Polska
[4]Centrum Badań Biokrystalograficznych, Instytut Chemii Bioorganicznej PAN, ul. Z. Noskowskiego 12/14, 61-704 Poznań, Polska

✉e-mail: wladek@iwonka.med.virginia.edu

*Pracę tę dedykujemy dr. Alexandrowi Wlodawerowi, niestrudzonemu ambasadorowi krystalografii białek

Słowa kluczowe: biologia strukturalna, struktura makromolekuł, Bank Struktur Białkowych (Protein Data Bank, PDB), bazy danych strukturalnych, walidacja struktury, czerpanie wiedzy z baz danych (data mining)

STRESZCZENIE

Bank Struktur Białkowych (Protein Data Bank, PDB), utworzony w 1971, gdy znano zaledwie 7 struktur białek, dziś zawiera ponad 120 tys. wyznaczonych doświadczalnie trójwymiarowych modeli makromolekuł biologicznych, w tym takich gigantów jak wirusy i rybosomy złożone z setek tysięcy atomów. Większość zdeponowanych w PDB struktur pochodzi z krystalografii rentgenowskiej, choć z postępem metodologii swój udział zaznaczyła spektroskopia NMR oraz – obecnie – kriomikroskopia elektronowa. Mimo że wyznaczanie struktury makromolekuł zostało z czasem ułatwione przez niewyobrażalny postęp techniki i metod obliczeniowych, jest ono nadal skomplikowanym procesem badawczym, wymagającym solidnego wyszkolenia, doświadczenia i talentu. Podobnie zrozumienie astronomicznej ilości danych zgromadzonych w PDB i przełożenie ich na wiedzę – wymaga dobrego przygotowania. Wstępem do niego może być niniejszy artykuł.