Determination of Protein Structures—A Series of Fortunate Events

Maksymilian Chruszcz,* Alexander Wlodawer,† and Wladek Minor*
*Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, Virginia; and †Protein Structure Section, Macromolecular Crystallography Laboratory, National Cancer Institute at Frederick, Frederick, Maryland

ABSTRACT Determination of a macromolecular structure using x-ray diffraction is a multistep process that involves a plethora of techniques involving molecular biology, bioinformatics, and physical sciences. Counterintuitively, the success of any or all individual steps does not guarantee the success of the overall process. This review examines the difficulties presented by each step on the path from a gene to the final publication, together with certain lucky (or unlucky) circumstances that can affect the velocity along that path.

INTRODUCTION

From its beginnings up to this day, protein crystallography (and, consequently, structural biology as a whole) owes a lot to fortunate events. The first such fortuitous circumstance was the lack of tenure system at the University of Cambridge, UK. This allowed Max Perutz to study the structure of hemoglobin for more than 20 years before the first significant structural results could be published (1). His work, however, allowed the development of a completely new methodology which was later used by all other groups investigating protein structures, and also led in 1962 to the award to Perutz of a Nobel Prize in chemistry. Moreover, the choice of hemoglobin as a target of that effort was a lucky one, as an unusually high fraction of the secondary structure of that protein is α-helical, which makes it very rigid, stable, well-diffracting, and comparatively easy to model.

Later, the development of synchrotrons by high-energy physicists catalyzed the explosion of protein structures solved by x-radiation. The orbiting particles, either electrons or positrons, generate what used to be called, in the early years, parasitic radiation. Thus, a small hole in the synchrotron wall could provide a source of x-radiation much stronger than any conventional generator. A number of dedicated synchrotron x-ray sources have been built all over the world since the early 1980s, followed by third-generation machines that generate x-rays not only by simple circulation of particles around the rings, but also by employing insertion devices called wigglers and undulators.

Thus, at present, crystallographers have access to more than 100 dedicated x-ray beamlines, located on 22 synchrotrons that have been constructed on all continents, except for Antarctica. How successful have they been? The synchrotron sources were responsible for a total of 3897 structures in 2005, more than three-quarters of all macromolecular structures published that year (Fig. 1). Some structures were solved in a matter of hours, if not minutes, after crystals were placed in the synchrotron beam (2). Obviously, however, only lucky events were reported, as on the average a single solved and deposited structure must have required roughly 40 h of synchrotron time (assuming 2000 h of synchrotron operation per year), as well as many priceless crystals.

The advancement in x-ray sources was accompanied by the development of fast x-ray detectors and by great advancement in computational methods and computer technology. All these developments also took place during the period of unprecedented progress in the techniques of molecular biology. Inexpensive workstations or even laptops have the computational power necessary to solve most crystallographic structures, and extremely sophisticated software is able to elucidate the three-dimensional structure even when very poorly diffracting crystals are used. Structure elucidation for a macromolecule is a multistep process (see Fig. S1 in Supplementary Material, Data S1) that requires 100% success at every step. A major difficulty in protein crystallography is that the success of a particular step can only be fully evaluated at the next step, or sometimes even two or three steps later. The experimenter may find it necessary to return to a previous (or possibly even the initial) step to achieve ultimate success. This iterative process of structure solution may take as long as 10–20 years of battle on various fronts. Some of our own projects that took very long to complete include, for example, the structure of nerve growth factor solved 17 years after crystals became available (3,4), or that of L-asparaginase, solved 19 years after initial crystallization (5,6). In some areas of structural genomics, however, where there is the option to drop a stubborn target, structure solution of a set of many targets can be compared to Napoleon's war with Russia. Napoleon did not lose a single battle, but lost the war of attrition, a term frequently used for description of the progress in structural genomics. The most productive centers were those that managed to win the war of attrition, not those that were the best in performing one or more steps (7,8). In this review, we will examine the difficulty presented by the various steps on the path from a gene to the final publication that should describe not only the structure, but also the mechanism

Submitted March 5, 2008, and accepted for publication March 31, 2008.

Address reprint requests to Wladek Minor, E-mail address: wladek@iwonka.med.virginia.edu.

Editor: Edward H. Egelman.

0006-3495/08/07/1/09 $2.00

doi: 10.1529/biophysj.108.131789

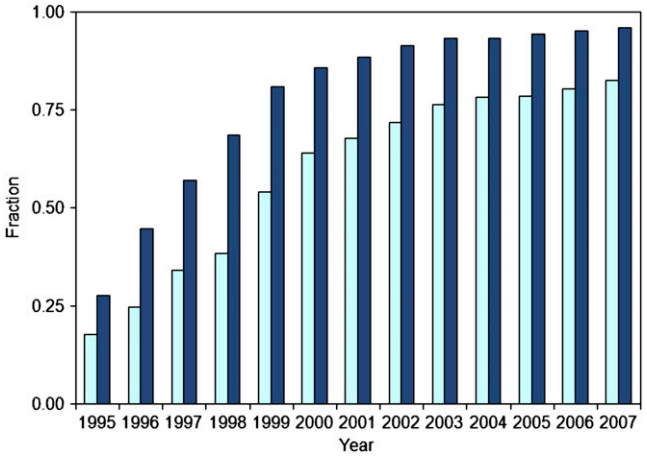

FIGURE 1 The fraction of x-ray structures deposited in the PDB that report the use of synchrotron sources for their determination (*dark blue*) and sample temperatures close to 100 K during data collection (*light blue*).

of action of a macromolecule and also the lucky (or unlucky) circumstances that can affect the velocity along that path.

From gene to crystal

Even in a high-throughput structural genomics center, the first step is a detailed analysis of all available data for a protein target, through bioinformatic and/or experimental approaches. Sometimes the experimental knowledge about a protein is nonexistent or very limited, but still there are many bioinformatic tools capable of extracting useful information, even when the gene sequence provides the only available information. For example, these tools often allow elimination of intrinsically unstructured proteins in the early stages of the process. It is generally agreed that work on mammalian and/or membrane proteins is much more challenging than on soluble bacterial proteins, but even a ''simple'' bacterial protein can cause a lot of trouble and may require substantial effort for structure solution. Even more challenging than studies of integral membrane proteins may be projects investigating the structure of various macromolecular complexes. Despite significant amounts of pure luck, solving such structures may take years. To give an example, ribosomal particles were crystallized as early as 1982 (9), but the first detailed structures were not completed until 2000 (10).

Cloning is a relatively straightforward step that is facilitated by a plethora of commercially available kits and services, including de novo gene synthesis. Synthetic genes with optimized codons (11) may improve the yield of protein, especially when they are expressed in a system utilizing different codon frequencies than their native genome. At first glance, the criterion for success of this step seems to be a simple one—either the gene is successfully cloned or not. Unfortunately, quite often problems encountered later, such as low expression level, lack of protein solubility, unsuccessful crystallization, or problematic properties of the crystals (e.g., twinning, low resolution diffraction) may result in the need to repeat the cloning step, despite its apparent success.

Constructs must be designed with the assumption that the protein itself should be treated as one of the most important factors that affect crystallization (12), and that structural analysis might require milligrams of protein. Why it is worth spending time choosing an appropriate expression system and vector? First of all, information about whether the protein folds properly by itself (13) or folds only under some special conditions (14,15) can influence vector design. For example, the target protein may require the presence of an additional protein which would act as a folding chaperone, protect it from degradation during expression, or assist in forming disulfide bonds (16). Crystallization experiments can take as long as weeks or months, so the protein should be stable for long periods of time. Proteins, especially those from eukaryotes, are frequently posttranslationally modified. In such cases, expression in a bacterial system may sometimes result in inactive protein, yet even that apparent failure could be advantageous for crystallization when, for example, the protein is not glycosylated (17). Proteins containing disulfide bonds (18) are especially difficult to express in the properly folded form when bacterial systems are used. Some proteins which are only expressed in inclusion bodies can be purified under denaturing conditions and then refolded, and consequently much attention has been paid to development of new refolding protocols (19). When refolding fails, the only choice is to select a different expression protocol (15) or expression system.

Sometimes the inability to purify and/or crystallize an intact protein may force the experimenter to switch tactics and choose a fragment (20) as a target for structure determination. Limited proteolysis is the technique of choice in determining the domain boundaries in multidomain proteins, but it is sometimes very difficult to choose a fragment of a protein that will be stable and represent a single domain. It can be difficult to make mammalian proteins in bacteria, and in many cases it is necessary to resort to the application of yeast, insect, or mammalian cells as expression systems. At all these stages we must think about the final goal of the experiment—elucidation of a three-dimensional structure that represents a biologically relevant form.

Once the yield of protein expression is deemed to be acceptable, the protein has to be purified. A hundred years ago, scientists used crystallization to purify proteins, as illustrated in a beautiful book showing hundreds of photographs of hemoglobin crystals (21). These days, it is assumed that to crystallize stubborn proteins, the sample must be homogeneous, not only in terms of the polypeptide sequence, but also in terms of protein folding, conformation, and possibly aggregation. To simplify protein purification and increase its speed, recombinant proteins are often fused with polypeptides or even full-size protein partners that facilitate affinity chromatography. By far the most popular purification

method (22) involves addition of a poly-histidine tag (His-tag) (23) in combination with metal-ion affinity chromatography (24). Such a tag consists of 6–10 histidine residues, usually followed by a spacer that allows subsequent cleavage by a suitable protease. One of the advantages of the His-tag is its relatively small size, meaning that sometimes the tag need not be removed before crystallization (25). Addition of fusion proteins often increases both the level of protein expression and its solubility. Fusing maltose-binding protein, thioredoxin, glutathione-S-transferase, or green fluorescent protein may be in many cases advantageous, and these proteins can be used alone or in combination with a His-tag (26). Moreover, the presence of a fusion partner, in particular maltose-binding protein, may help to properly fold the passenger protein (27).

Crystallization requires samples of a protein in milligram quantities, although with sufficient luck, combined with new nanotechnologies, the amount of protein required to find the initial crystallization conditions might be significantly reduced (28). Unfortunately, nanoliter technologies often produce crystals that are too small for x-ray structure analysis. The transfer of nanoliter crystallization conditions to the microliter scale is not always straightforward and the nature of the difficulties in scaleup is not fully understood (although development of in-chip techniques (29) may change that situation). Initial screening is most often performed using the sparse matrix method (30), and a variety of commercial screens optimized for the crystallization of proteins (31,32), nucleic acids, macromolecular complexes (33), or membrane proteins are currently available. If the experimenter is lucky, after setting up several hundred crystallization conditions she or he may start optimization of crystal growth. The process of crystal optimization can be performed in many ways, and in most common approaches, grid screen designs (34) based on the initially obtained conditions are used. Other approaches involve addition of so-called additives—usually small-molecule compounds—to the crystallization media (35). Less fortunate experimenters who failed in finding any conditions for crystal growth may try reductive methylation of lysine residues (36) before they return to designing new protein constructs. In cases where reductive methylation fails, sequence mutation(s) could be the next choice for increasing crystallizability of a protein (22,37), but such an approach requires return to the beginning of the path. An alternative rescue strategy is the use of in situ proteolysis (38).

Twenty years ago, once a crystal was grown, it was placed in a sealed capillary in the presence of its mother liquor and was used for diffraction experiments conducted at room temperature on a laboratory-based x-ray source. Now, however, a vast majority of the diffraction experiments are performed at synchrotrons (Fig. 1). The flux at some high-intensity beamlines is so high that an unprotected crystal would evaporate in milliseconds. Even relatively low-intensity synchrotron beams induce radiation damage (39), and may cause a variety of chemical modifications of the protein (Fig. 2). The most efficient way of slowing down that process is cryocooling (40), which, in connection with a very simple method of crystal mounting using the so-called cryo-loops (41), has revolutionized data collection (Fig. 1). During cryocooling, crystals protected by cryosolutions are rapidly transferred to a nitrogen stream maintained at a temperature near 100 K. Under such conditions, the solution around and inside the crystal is glassified. The cryosolutions may contain different types of alcohols, salts, or oils that prevent ice formation which would destroy the order of macromolecular crystals. Crystal freezing, although relatively simple, requires testing of several different cryosolutions, but even exhaustive cryocooling experiments may produce samples of much lower quality than the original crystals. Altering the crystal environment, especially by controlling the humidity or by annealing, may dramatically improve crystal quality. Recently, there have been many reports about the use of a credit-card approach, although no financial transactions are involved. A piece of thin plastic (a credit card would do quite well) can be used to block the cryostream for several seconds and anneal the crystal (42). Annealing may lead to dramatic improvement of diffraction quality (see Fig. S2 in Data S1), but such a result is by no means guaranteed. Thus, we would

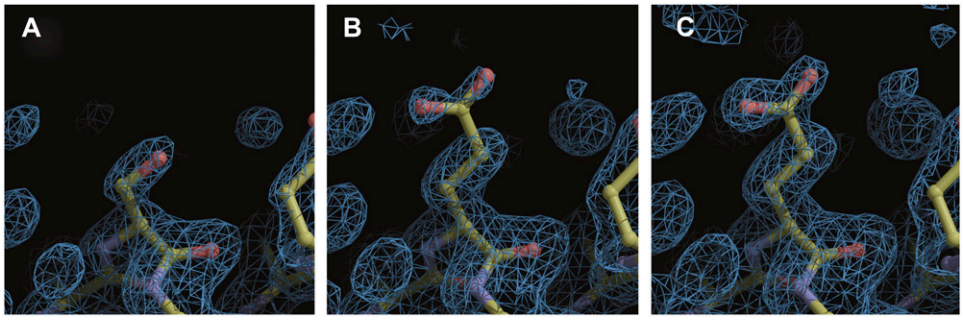

FIGURE 2 Residue 160 in the crystal structure of soybean lipoxygenase shown together with $2F_o$-F_c electron density map. (*A*) A map contoured at 1σ level. The crystal was obtained using natural protein isolated from soybeans and the residue was identified as serine (PDB codes: 1YGE and 1F8N). (*B*) Map shown at 0.7σ. The serine was replaced by glutamic acid, what is consistent with results of the DNA sequencing (Ted Holman, 2007, private communication). (*C*) Electron density shown at 0.3σ. Different contouring of the map reveals either possible decarboxylation during data collection or conformational flexibility of Glu^{160}.

not like to recommend such an approach in cases where data quality is questionable, but only one crystal is available, and the experimenter has doubts whether the crystallography gods will smile on him/her on that particular day.

From diffraction images to electron density map

Placement of a crystal in the x-ray beam initiates the last experimental step of the process of structure determination. In principle, this experiment is a very simple one, so it is believed that data collection should be easy and straightforward to automate, as there are only a very few parameters that are under control of the experimenter. These parameters include crystal/detector distance, exposure time, oscillation angle, and wavelength of x-radiation. However, there are at least three additional parameters that are beyond the experimenter's control: crystal quality (long-range order, mosaic spread), radiation decay, and limitations of the experimental setup (e.g., detector dynamic range, goniostat precision, etc.). The difficulty of choosing user-controlled parameters that minimize the detrimental influence of crystal quality and radiation decay is illustrated by analysis of the data collected on one of the ALS beamlines. Data from this beamline show that, on average, it is necessary to collect 57 full data sets (43) to make one PDB deposit, and the number of tested crystals is even higher. The experimental difficulty lays in the fact that the result of a diffraction experiment is a set of diffraction intensities (or amplitudes), not the phases that are necessary for calculation of the electron density map.

In the current practice, diffraction data are collected for three major types of calculations: molecular replacement (MR), multiple anomalous diffraction (MAD)/single anomalous diffraction (SAD), and the final refinement of the model. The previously popular multiple isomorphous replacement has been overshadowed by the use of techniques based on anomalous scattering. In the case of MR experiments (44), the source of phases is a model of the same or a similar protein, and the accuracy of the measured intensities is much less important than obtaining a complete set, without the loss of strong peaks through oversaturation of the detector. For solution of new structures by the SAD, MAD, or even multiple isomorphous replacement techniques, the phases are derived from the differences between the observed diffraction intensities and thus their accuracy is of utmost importance. Data collection for the final model refinement has a simple goal—to collect complete, high-accuracy data to the resolution limit of diffraction. The latter experiment seems to be the easiest one, but even it requires careful planning, as improvement of statistical accuracy of measured intensities does not necessarily result in better data (longer counting time may increase radiation damage).

Traditionally, SAD/MAD experiments which require highly accurate data were considered to be particularly difficult, but development of experimental hardware, software, and protocols has increased substantially the percentage of structures solved by these techniques. Since a SAD experiment involves collecting only a fraction of data required for MAD, it should be expected that the former method would be preferable, but the differences between various regions of the world in the use of these techniques show that proliferation of the most efficient experimental protocols is slow (Fig. 3). SAD/MAD experiments rarely fail just because of the lack of a sufficient number of atoms that produce anomalous signal (except when very weak anomalous scatterers such as sulfur are used), but more often due to experimental errors, involving too many saturated detector pixels (overloads), or an improper data collection strategy that may result in premature radiation damage and/or incomplete data. Even small errors at this stage mean significantly more work for the experimenter during structure solution and refinement. In many cases, less than optimal diffraction experiments have to be repeated. The completeness of low resolution data is quite often neglected (PDB deposits report completeness in the highest resolution shell but not in the lowest), resulting in difficulties during structure solution and model building. Sometimes, given pure luck and experience, one may still recover from such problems using nonstandard approaches (45), but it has to be stressed that accurate, nonsaturated low-resolution data are critical for both the MR and SAD/MAD techniques. It is sometimes a surprise that structure can be solved using data from a lower quality crystal, rather than from a set collected on a high quality crystal. Low quality crystals diffract weakly and consequently do not produce overloaded low resolution reflections. In that case the order in which crystals are chosen for the diffraction experiments may determine the probability of the final success, as rarely do experimenters collect another complete data set when they have already seen ''perfect'' diffraction.

In the last several years, interpretation of experimental results has been greatly facilitated by several integrated software packages such as CCP4 (46), PHENIX (47), or HKL-3000 (48), coupled with the availability of fast com-

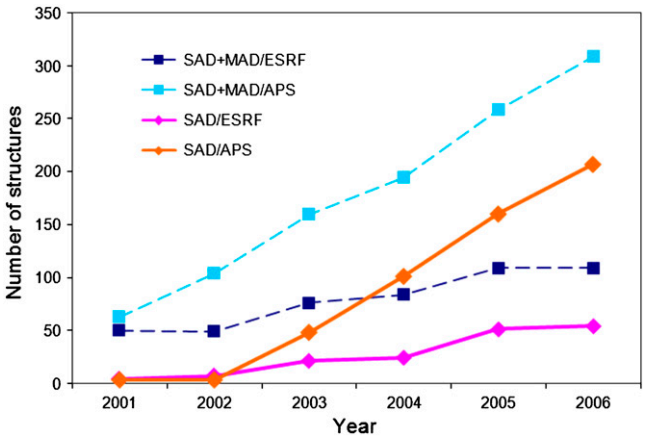

FIGURE 3 Comparison of the extent of application of SAD and MAD techniques in APS and ESRF, as reported in PDB.

puters that allow for almost real-time, on-the-fly calculations. The methods applied to solution of the phase problem, although quite complicated, are hidden behind sophisticated software and (sometimes even more sophisticated) user interfaces that can occasionally make a noncrystallographer a competent and efficient structure solver. In simple cases, an interpretable initial electron density map may be obtained with just a few clicks of a mouse.

Most often, the initial phases (and, as a consequence, electron density maps) obtained in SAD/MAD or MR experiments are not very accurate, and their interpretation could be very difficult. Luckily, several phase improvements methods have been developed that, when properly applied, may dramatically increase the quality of the electron density maps. Solvent-flattening methods and noncrystallographic symmetry averaging are especially popular. It is worth mentioning that although crystals with high solvent content often diffract poorly (49), nevertheless the high solvent content may turn out to be advantageous and help produce a high quality initial map. Dramatic changes in the quality of electron density maps can be observed when noncrystallographic symmetry averaging is applied, so the presence of many (but not too many) copies of a macromolecule in an asymmetric unit should not be considered bad luck.

Even in the cases of properly performed data collection, the nature of the crystals may introduce problems that prevent structure solution. One of the most vexing problems is caused by twinning, a phenomenon that arises when more than one lattice diffracts simultaneously. It was recently reported (50) that the combination of crystal and lattice symmetries could allow twinning in more than 30% of cases of the structures reported in PDB. Moreover, twinning is not always noticed, and in some situations prevents structure solution. If that happens, there is no other choice than to return to the laboratory and grow a new crystal form of a particular macromolecule.

Model building, refinement, and structure validation

In the current practice not only are the initial electron density maps generated in an automatic or semiautomatic manner by software, but also the interpretation of the resulting electron density may be done almost automatically. Several programs, including ARP/wARP (51), RESOLVE (52), and MAIN (53), use different approaches to automated model building. Manual model building and adjustment also become relatively easy, thanks to powerful graphics software such as O (54) and COOT (55). However, it is still not trivial to interpret the electron density maps obtained at low resolution (especially below 3.2 Å).

Final refinement of macromolecular structures is usually accomplished with programs such as CNS/CNX/X-PLOR (56), REFMAC (57), and SHELXL (58). Structure refinement is comparatively easy if data extend to between 1.5 and 2.4 Å. Refinement with very high resolution data can be time-consuming due to the wealth of structural details that have to be modeled (such as multiple conformations of the side chains, complicated temperature factor models, etc.). Very low resolution structures (below 3.2 Å) are in a separate class, requiring very careful refinement and validation. Another type of difficulty in refinement arises when a structure contains moieties other than amino acids, such as metal ions and small molecule compounds. Although identification and refinement of metal ions present in protein structure seems to be comparatively straightforward (59), many new structures reported in the PDB still contain metal ions with very improbable coordination or geometry of the metal-binding environment. Protein-DNA or protein-ligand complexes may pose additional difficulties as automatic model building works best for amino-acid chains. A refinement strategy depends mostly on resolution, and the resolution also determines how many parameters may be refined, and how they should be treated (60).

Refinement and manual structure rebuilding or adjustment has to be performed together with model validation. As mentioned above, significant advances in software allow many noncrystallographers to collect data, as well as solve and refine x-ray structures, without advanced knowledge of the underlying techniques. Especially in such cases, sophisticated tools for structure validation are necessary. Validation tools should not only detect serious crystallographic and chemical errors in the models, but should also be able to guide an inexperienced person and suggest how to correct errors. Examples of such programs are PROCHECK (61), WHATCHECK (62), MOLPROBITY, and KING (63). Sometimes experimenters ignore clear warnings from the validation programs even during the process of deposition in PDB, presumably since they are convinced that their structure is so special that any violations of known chemical rules simply support its uniqueness. Unfortunately, only a very small percent of lucky scientists observing novel chemistry in their structures will ultimately hear from the Nobel Committee, but the unlucky ones will find sooner or later that the validation tools will ruin their claims. Since deposition of structure factors is now required in practically all publicly funded research, other crystallographers can now routinely use the ultimate validation tool, i.e., re-evaluation of questionable structures, so it is unlikely that the wrongly refined structures will be able to pollute the databases in the future.

Interpretation of a model

It should be always remembered that the ultimate aim of a crystallographic experiment, even if conducted under the umbrella of structural genomics, is not creation of just a model consisting of the atomic coordinates, but rather providing guidance to interpretation of chemical and biological information. However, interpretation of the models should be done in a way that takes into consideration their limitations

imposed by factors such as, for example, data resolution, overall quality of the model as indicated by R/R_{free}, as well as its chemical correctness. Moreover, it is worth stressing that the final model does not represent a single molecule, but is a time and space average from many molecules. High energy radiation, particularly originating from very bright synchrotron beamlines, is able to cause chemical modification of molecules, and, as shown in Fig. 2, even a model obtained from high resolution data cannot be treated as error-free. At low resolution, misinterpretation of the electron density is relatively easy, and a careless approach to such data may result in tracing a fragment of the amino-acid chain in the opposite direction, or, very rarely, producing a completely incorrect model. Moreover, the representation of numeric values in the atomic coordinates deposited in the PDB format (three digits after the decimal point) may be very misleading to an inexperienced experimenter, who may assume that all digits are significant and analyze the structure according to that assumption (64–66). Analysis of structures deposited in the PDB should take into account that the models contain different types of errors which accumulated during the whole process of structure determination, so interpretation of a three-dimensional structure and all chemical or biological conclusions derived from it are strongly affected by the quality of the model. In particular, deduction of a detailed mechanism of an enzymatic reaction requires knowledge of the hydrogen-bond network in the macromolecule of interest. Unfortunately, only the luckiest experimenters who are able to determine a protein structure at very high resolution may directly observe hydrogens in their structures. For most structures (60% of the structures in PDB are between 1.7 and 2.5 Å resolution), the interpretation is not direct and a single structure may support multiple chemical or biological reaction mechanisms. A similar problem has to be solved by a translator of poetry. Translation is an art, and a poem translated into a new language may even be better than the original. Similarly, reinterpretation of a structure is quite often much better than the original. The process of moving from the coordinates to interpretation of the mechanism of action is the most difficult step.

Is the structure biologically relevant?

Once the structure has been solved at high resolution, with low R factors and small departure of the geometric parameters from the library values, how confident can we be that it describes a biologically relevant state of the protein? That question has been asked (and answered) many times since the beginning of protein crystallography. It is actually not a single problem, but at least two interrelated ones. The first, and maybe the easiest one to answer, is the question of whether the structure of a protein in the crystal (solid state) is the same as in solution. An early example was provided by careful analysis of the structures of a small helical cytokine, interleukin-4, solved independently in four laboratories. Two structures of this protein were obtained by crystallography, and two other by NMR. Their comparison has clearly shown that the differences between these structures were due more to the uncertainties in their determination (much larger for NMR than crystallography) than to any variations in the proteins (67). Thus, although this is a legitimate concern and still needs to be answered separately in each specific case, a general answer is that usually the differences between solution and solid state of proteins are small, if any.

Another part of the question, though, deals with the relevance of the observed structure for the explanation of the biological properties of the molecular system under study, and it does not have a unique answer. Let us consider an enzyme and the details of the reaction that it catalyzes. Clearly, the structure of the apoenzyme may not be sufficient to describe all steps of the reaction, since parts of the active site may adjust to the presence of the substrate, transition state, and product, and the nature of such changes is not always easy to predict. In particular, the structure of the transition state would be most illuminating, but by definition it is not directly accessible, since it is unstable on the crystallographic timescale. Utilization of transition state mimics and extremely fast data collection using Laue crystallography (68) can help, but they still do not provide a guarantee that the state of the protein observed in the crystal can directly explain its biologically relevant properties, since proteins are by no means stationary.

A relevant example of a plethora of difficulties encountered in determining the biological properties of a protein based on crystallographic investigations is provided by the ATP-dependent protease Lon. Although the enzyme has been known for more than 20 years and its crystallographic investigations span a decade, the full-length Lon has resisted crystallization (69). However, since the domain structure of Lon has been determined, its individual domains have been crystallized and analyzed separately. This work yielded a number of surprises. For example, the structure of the active site was significantly different in the catalytic domain of Lon isolated from different bacterial sources, and it was initially suggested that these differences might play a biological role (70). However, subsequent crystallographic and mutagenesis studies yielded a rather different picture, suggesting that none of the structures of the apoenzyme show the active site in a biologically relevant state, as the presence of a substrate or a product of the reaction is likely to reorganize it very significantly. Since no good and specific substrates of Lon are known, the details of its mechanism of action are still not understood, even though an atomic-resolution structure of its catalytic domain is available (71). Even more difficult is analysis of the biological properties of the N-terminal domain of this protein, most likely involved in substrate binding. Crystal structure of the construct containing just over 100 residues indicated a novel fold, not found in any known protein complexes and thus no conclusions about the mode of binding could be drawn (72). However, a structure of a

hypothetical protein from *Bordetella parapertussis*, BPP1347, deposited in the PDB by the Northeast Structural Genomics Consortium, has shown very significant topological similarity, despite very low sequence similarity. This example illustrates a common problem with some structural genomics-derived structures, namely a difficulty of assigning the function to proteins with novel fold (and that, incidentally, is one of the stated reasons for these undertakings), but even structures obtained in targeted efforts may not fare much better.

SUMMARY

In structural biology, the path from gene to publication most often requires a significant amount of work and much luck. In many cases, the period between obtaining the initial crystallization conditions and publishing a structure may extend over a decade. The bottlenecks of the whole process as well as the frequently-used term, ''high-hanging fruit,'' are being constantly redefined. The major difficulty is the same as in any other cutting-edge experimental science—the extraction of a low signal from high noise. Nowadays, the whole process of determining macromolecular structures is faster then ever (see Fig. S3 in Data S1). We predict that, although the proliferation of the best experimental protocols and the development of new methodologies will decrease crystallographers' dependence on fortunate circumstances, luck will still play a significant role in the foreseeable future.

SUPPLEMENTARY MATERIAL

To view all of the supplemental files associated with this article, visit www.biophysj.org.

We thank Dominika Borek, Zbyszek Dauter, Aled Edwards, Andrzej Joachimiak, Zbyszek Otwinowski, and Matthew Zimmerman for valuable comments.

W.M. was supported by grants No. GM74942 and No. GM53163 and the original work in the laboratory of A.W. was supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research.

REFERENCES

1. Perutz, M. F., M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. T. North. 1960. Structure of hemoglobin. A three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by x-ray analysis. *Nature*. 185:416–421.
2. Cassetta, A., A. M. Deacon, S. E. Ealick, J. R. Helliwell, and A. W. Thompson. 1999. Development of instrumentation and methods for MAD and structural genomics at the SRS, ESRF, CHESS and Elettra facilities. *J. Synchr. Rad.* 6:822–833.
3. Wlodawer, A., K. O. Hodgson, and E. M. Shooter. 1975. Crystallization of nerve growth factor from mouse submaxillary glands. *Proc. Natl. Acad. Sci. USA*. 72:777–779.
4. McDonald, N. Q., R. Lapatto, J. Murray-Rust, J. Gunning, A. Wlodawer, and T. L. Blundell. 1991. New protein fold revealed by a 2.3-Å resolution crystal structure of nerve growth factor. *Nature*. 354: 411–414.
5. Wlodawer, A., K. O. Hodgson, and K. Bensch. 1975. Studies of two crystal forms of L-glutaminase-asparaginase from *Acinetobacter glutaminasificans*. *J. Mol. Biol.* 99:295–299.
6. Swain, A. L., M. Jaskólski, D. Housset, J. K. Rao, and A. Wlodawer. 1993. Crystal structure of *Escherichia coli* L-asparaginase, an enzyme used in cancer therapy. *Proc. Natl. Acad. Sci. USA*. 90:1474–1478.
7. Chandonia, J. M., and S. E. Brenner. 2006. The impact of structural genomics: expectations and outcomes. *Science*. 311:347–351.
8. O'Toole, N., M. Grabowski, Z. Otwinowski, W. Minor, and M. Cygler. 2004. The structural genomics experimental pipeline: insights from global target lists. *Proteins*. 56:201–210.
9. Yonath, A., J. Mussig, and H. G. Wittmann. 1982. Parameters for crystal growth of ribosomal subunits. *J. Cell. Biochem.* 19:145–155.
10. Ban, N., P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*. 289:905–920.
11. Hoover, D. M., and J. Lubkowski. 2002. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* 30:e43.
12. Dale, G. E., C. Oefner, and A. D'Arcy. 2003. The protein as a variable in protein crystallization. *J. Struct. Biol.* 142:88–97.
13. Dunker, A. K., Z. Obradovic, P. Romero, E. C. Garner, and C. J. Brown. 2000. Intrinsic protein disorder in complete genomes. *Genome Inform.* 11:161–171.
14. Bourhis, J. M., B. Canard, and S. Longhi. 2007. Predicting protein disorder and induced folding: from theoretical principles to practical applications. *Curr. Protein Pept. Sci.* 8:135–149.
15. Oganesyan, N., I. Ankoudinova, S. H. Kim, and R. Kim. 2007. Effect of osmotic stress and heat shock in recombinant protein overexpression and crystallization. *Protein Expr. Purif.* 52:280–285.
16. Stols, L., M. Zhou, W. H. Eschenfeldt, C. S. Millard, J. Abdullah, F. R. Collart, Y. Kim, and M. I. Donnelly. 2007. New vectors for co-expression of proteins: structure of *Bacillus subtilis* ScoAB obtained by high-throughput protocols. *Protein Expr. Purif.* 53:396–403.
17. Chang, V. T., M. Crispin, A. R. Aricescu, D. J. Harvey, J. E. Nettleship, J. A. Fennelly, C. Yu, K. S. Boles, E. J. Evans, D. I. Stuart, R. A. Dwek, E. Y. Jones, R. J. Owens, and S. J. Davis. 2007. Glycoprotein structural genomics: solving the glycosylation problem. *Structure*. 15:267–273.
18. Woycechowsky, K. J., B. A. Hook, and R. T. Raines. 2003. Catalysis of protein folding by an immobilized small-molecule dithiol. *Biotechnol. Prog.* 19:1307–1314.
19. Kaulmann, G., G. J. Palm, K. Schilling, R. Hilgenfeld, and B. Wiederanders. 2003. An unfolding/refolding step helps in the crystallization of a poorly soluble protein. *Acta Crystallogr. D*. 59:1243–1245.
20. Malawski, G. A., R. C. Hillig, F. Monteclaro, U. Eberspaecher, A. A. Schmitz, K. Crusius, M. Huber, U. Egner, P. Donner, and B. Muller-Tiemann. 2006. Identifying protein construct variants with increased crystallization propensity—a case study. *Protein Sci.* 15:2718–2728.
21. Reichert, E. T., and A. P. Brown. 1909. The differentiation and specificity of corresponding proteins and other vital substances in relation to biological classification and organic evolution. The crystallography of hemoglobins. The Carnegie Institution of Washington, Publication No. 116.
22. Derewenda, Z. S. 2004. The use of recombinant methods and molecular engineering in protein crystallization. *Methods*. 34:354–363.
23. Smith, M. C., T. C. Furman, T. D. Ingolia, and C. Pidgeon. 1988. Chelating peptide-immobilized metal ion affinity chromatography. A new concept in affinity chromatography for recombinant proteins. *J. Biol. Chem.* 263:7211–7215.
24. Porath, J., J. Carlsson, I. Olsson, and G. Belfrage. 1975. Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature*. 258:598–599.
25. Carson, M., D. H. Johnson, H. McDonald, C. Brouillette, and L. J. Delucas. 2007. His-tag impact on structure. *Acta Crystallogr. D*. 63: 295–301.

26. Donnelly, M. I., M. Zhou, C. S. Millard, S. Clancy, L. Stols, W. H. Eschenfeldt, F. R. Collart, and A. Joachimiak. 2006. An expression vector tailored for large-scale, high-throughput purification of recombinant proteins. *Protein Expr. Purif.* 47:446–454.

27. Sachdev, D., and J. M. Chirgwin. 1998. Solubility of proteins isolated from inclusion bodies is enhanced by fusion to maltose-binding protein or thioredoxin. *Protein Expr. Purif.* 12:122–132.

28. Li, L., D. Mustafi, Q. Fu, V. Tereshko, D. L. Chen, J. D. Tice, and R. F. Ismagilov. 2006. Nanoliter microfluidic hybrid method for simultaneous screening and optimization validated with crystallization of membrane proteins. *Proc. Natl. Acad. Sci. USA.* 103:19243–19248.

29. Anderson, M. J., B. DeLabarre, A. Raghunathan, B. O. Palsson, A. T. Brunger, and S. R. Quake. 2007. Crystal structure of a hyperactive *Escherichia coli* glycerol kinase mutant $Gly^{230} \rightarrow Asp$ obtained using microfluidic crystallization devices. *Biochemistry.* 46:5722–5731.

30. Jancarik, J., and S. H. Kim. 1991. Sparse-matrix sampling—a screening method for crystallization of proteins. *J. Appl. Cryst.* 24:409–411.

31. Wooh, J. W., R. D. Kidd, J. L. Martin, and B. Kobe. 2003. Comparison of three commercial sparse-matrix crystallization screens. *Acta Crystallogr. D.* 59:769–772.

32. Newman, J., D. Egan, T. S. Walter, R. Meged, I. Berry, M. Ben Jelloul, J. L. Sussman, D. I. Stuart, and A. Perrakis. 2005. Towards rationalization of crystallization screening for small- to medium-sized academic laboratories: the PACT/JCSG+ strategy. *Acta Crystallogr. D.* 61:1426–1431.

33. Radaev, S., S. Li, and P. D. Sun. 2006. A survey of protein-protein complex crystallizations. *Acta Crystallogr. D.* 62:605–612.

34. Carter, C. W., Jr., and Y. Yin. 1994. Quantitative analysis in the characterization and optimization of protein crystal growth. *Acta Crystallogr. D.* 50:572–590.

35. McPherson, A., and B. Cudney. 2006. Searching for silver bullets: an alternative strategy for crystallizing macromolecules. *J. Struct. Biol.* 156:387–406.

36. Walter, T. S., C. Meier, R. Assenberg, K. F. Au, J. Ren, A. Verma, J. E. Nettleship, R. J. Owens, D. I. Stuart, and J. M. Grimes. 2006. Lysine methylation as a routine rescue strategy for protein crystallization. *Structure.* 14:1617–1622.

37. Derewenda, Z. S., and P. G. Vekilov. 2006. Entropy and surface engineering in protein crystallization. *Acta Crystallogr. D.* 62:116–124.

38. Dong, A., X. Xu, A. M. Edwards, C. Chang, M. Chruszcz, M. Cuff, M. Cymborowski, R. Di Leo, O. Egorova, E. Evdokimova, E. Filippova, J. Gu, J. Guthrie, A. Ignatchenko, A. Joachimiak, N. Klostermann, Y. Kim, Y. Korniyenko, W. Minor, Q. Que, A. Savchenko, T. Skarina, K. Tan, A. Yakunin, A. Yee, V. Yim, R. Zhang, H. Zheng, M. Akutsu, C. Arrowsmith, G. V. Avvakumov, A. Bochkarev, L. G. Dahlgren, S. Dhe-Paganon, S. Dimov, L. Dombrovski, P. Finerty, Jr., S. Flodin, A. Flores, S. Graslund, M. Hammerstrom, M. D. Herman, B. S. Hong, R. Hui, I. Johansson, Y. Liu, M. Nilsson, L. Nedyalkova, P. Nordlund, T. Nyman, J. Min, H. Ouyang, H. W. Park, C. Qi, W. Rabeh, L. Shen, Y. Shen, D. Sukumard, W. Tempel, Y. Tong, L. Tresagues, M. Vedadi, J. R. Walker, J. Weigelt, M. Welin, H. Wu, T. Xiao, H. Zeng, and H. Zhu. 2007. In situ proteolysis for protein crystallization and structure determination. *Nat. Methods.* 4:1019–1021.

39. Hendrickson, W. A. 1976. Radiation damage in protein crystallography. *J. Mol. Biol.* 106:889–893.

40. Garman, E., and R. L. Owen. 2006. Cryocrystallography of macromolecules: practice and optimization. *Methods Mol. Biol.* 364:1–18.

41. Teng, T. Y. 1990. Mounting of crystals for macromolecular crystallography in a freestanding thin-film. *J. Appl. Cryst.* 23:387–391.

42. Hanson, B. L., J. M. Harp, and G. J. Bunick. 2003. The well-tempered protein crystal: annealing macromolecular crystals. *Methods Enzymol.* 368:217–235.

43. Holton, J. 2005. Teaching elves to collect data: an analysis of the last million diffraction images from ALS 8.3.1. *In* Annual Meeting of the American Crystallographic Association, Orlando, FL.

44. Rossmann, M. G., and D. M. Blow. 1962. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr.* 15:24–31.

45. Koclega, K. D., M. Chruszcz, M. D. Zimmerman, M. Cymborowski, E. Evdokimova, and W. Minor. 2007. Crystal structure of a transcriptional regulator TM1030 from *Thermotoga maritima* solved by an unusual MAD experiment. *J. Struct. Biol.* 159:424–432.

46. Collaborative Computational Project Number 4. 1994. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D.* 50:760–763.

47. Adams, P. D., R. W. Grosse-Kunstleve, L. W. Hung, T. R. Ioerger, A. J. McCoy, N. W. Moriarty, R. J. Read, J. C. Sacchettini, N. K. Sauter, and T. C. Terwilliger. 2002. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D.* 58:1948–1954.

48. Minor, W., M. Cymborowski, Z. Otwinowski, and M. Chruszcz. 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.

49. Kantardjieff, K. A., and B. Rupp. 2003. Matthews coefficient probabilities: improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci.* 12:1865–1871.

50. Lebedev, A. A., A. A. Vagin, and G. N. Murshudov. 2006. Intensity statistics in twinned crystals with examples from the PDB. *Acta Crystallogr. D.* 62:83–95.

51. Perrakis, A., R. Morris, and V. S. Lamzin. 1999. Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.* 6:458–463.

52. Terwilliger, T. 2004. SOLVE and RESOLVE: automated structure solution, density modification and model building. *J. Synchr. Rad.* 11: 49–52.

53. Turk, D. 2001. Towards automatic macromolecular crystal structure determination. *In* Methods in Macromolecular Crystallography, NATO Science Series I. D. Turk and L. Johnson, editors. IOS Press, Amsterdam, The Netherlands.

54. Jones, T. A., J. Y. Zou, S. W. Cowan, and M. Kjeldgaard. 1991. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. D.* 47:110–119.

55. Emsley, P., and K. Cowtan. 2004. COOT: model-building tools for molecular graphics. *Acta Crystallogr. D.* 60:2126–2132.

56. Brünger, A. T., P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren. 1998. Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D.* 54:905–921.

57. Murshudov, G. N., A. A. Vagin, and E. J. Dodson. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D.* 53:240–255.

58. Sheldrick, G. M., and T. R. Schneider. 1997. SHELXL: high-resolution refinement. *Macromol. Crystallogr. Meth. Enzymol. B.* 277:319–343.

59. Harding, M. M. 2004. The architecture of metal coordination groups in proteins. *Acta Crystallogr. D.* 60:849–859.

60. Jaskólski, M., M. Gilski, Z. Dauter, and A. Wlodawer. 2007. Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? *Acta Crystallogr. D.* 63:611–620.

61. Laskowski, R. A., M. W. Macarthur, D. S. Moss, and J. M. Thornton. 1993. PROCHECK—a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* 26:283–291.

62. Hooft, R. W., G. Vriend, C. Sander, and E. E. Abola. 1996. Errors in protein structures. *Nature.* 381:272.

63. Lovell, S. C., I. W. Davis, W. B. Arendall 3rd, P. I. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. 2003. Structure validation by Cα geometry: ϕ, ψ, and Cβ deviation. *Proteins.* 50:437–450.

64. Wlodawer, A. 2007. Deposition of structural data redux. *Acta Crystallogr. D.* 63:421–423.

65. Wlodawer, A., W. Minor, Z. Dauter, and M. Jaskólski. 2008. Protein crystallography for non-crystallographers, or how to get the best (but

not more) from published macromolecular structures. *FEBS J.* 275: 1–21.

66. Dauter, Z., and T. Baker. 2007. Numerology. *Acta Crystallogr. D.* 63:275.
67. Smith, L. J., C. Redfield, R. A. Smith, C. M. Dobson, G. M. Clore, A. M. Gronenborn, M. R. Walter, T. L. Naganbushan, and A. Wlodawer. 1994. Comparison of four independently determined structures of human recombinant interleukin-4. *Nat. Struct. Biol.* 1:301–310.
68. Hajdu, J., K. R. Acharya, D. I. Stuart, P. J. McLaughlin, D. Barford, N. G. Oikonomakos, H. Klein, and L. N. Johnson. 1987. Catalysis in the crystal: synchrotron radiation studies with glycogen phosphorylase *b*. *EMBO J.* 6:539–546.
69. Rotanova, T. V., I. Botos, E. E. Melnikov, F. Rasulova, A. Gustchina, M. R. Maurizi, and A. Wlodawer. 2006. Slicing a protease: structural features of the ATP-dependent Lon proteases gleaned from investigations of isolated domains. *Protein Sci.* 15:1815–1828.
70. Im, Y. J., Y. Na, G. B. Kang, S. H. Rho, M. K. Kim, J. H. Lee, C. H. Chung, and S. H. Eom. 2004. The active site of a Lon protease from *Methanococcus jannaschii* distinctly differs from the canonical catalytic dyad of Lon proteases. *J. Biol. Chem.* 279:53451–53457.
71. Botos, I., E. E. Melnikov, S. Cherry, S. Kozlov, O. V. Makhovskaya, J. E. Tropea, A. Gustchina, T. V. Rotanova, and A. Wlodawer. 2005. Atomic-resolution crystal structure of the proteolytic domain of *Archaeoglobus fulgidus* Lon reveals the conformational variability in the active sites of lon proteases. *J. Mol. Biol.* 351:144–157.
72. Li, M., F. Rasulova, E. E. Melnikov, T. V. Rotanova, A. Gustchina, M. R. Maurizi, and A. Wlodawer. 2005. Crystal structure of the N-terminal domain of *E. coli* Lon protease. *Protein Sci.* 14:2895–2900.