PREFACE

A LOOK AT THE FUTURE OF MACROMOLECULAR STRUCTURE DETERMINATION

DUILIO CASCIO, KENNETH GOODWILL AND EDWARD MARCOTTE

UCLA-DOE Laboratory of Structural Biology, Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095

One of the reasons we love science is its predictive power, but can we predict the future of science itself? Here, we'll gaze into our crystal ball at the future of just one discipline, macromolecular crystallography, and make some guesses as to how the discipline will evolve during our lifetimes. In the past few decades, macromolecular crystallographers have improved and streamlined all of the steps of solving structures-from cloning, purification, and crystallization, to phasing and refinement. But even today, a well-established lab with experienced molecular biologists and crystallographers will often take six months to a year to do all these steps. Why can't structures be solved more quickly, as in the case of small molecule crystallography? What are the bottlenecks, and which techniques need to be improved?

Let's review the steps that brought macromolecular crystallography from an arcane art to the verge of a mass production technology.

The 70's

A crystallographer obtains purified protein from a collaborator. With minimal in-house handling and a bit of luck, the protein crystallizes after a long, agonizing year. Solving the structure then requires the following regime of time-consuming steps: Characterize the crystals by precession photography. To find heavy atom derivatives, collect several hundred oscillation photos from dozens of crystals mounted in capillaries at room temperature. Scan the photos at home in an Optronics or Jeol scanner. Solve the Patterson manually and calculate the phases with homemade programs. Build the model in a Richard's box. The total time to complete the project from protein sample to a fully refined molecule is around 5 to 7 years.

The 80's

Crystallographers grow tired of waiting for the 10 mg of protein from the collaborators, so they start purifying the protein in-house. Data collection is much faster with multiwire area detectors; now it is possible to screen heavy atom derivatives in a few days. Large groups travel to the synchrotron to collect data on Nonius Fast systems or on Fuji imaging plates and offline scanners. Still, the great majority of crystals are collected at room temperature. The total time to complete the project from protein sample to a fully refined molecule is around 2 to 5 years

From 1990 to 1995

We see the first complete in-house cloning, expression, purification, and crystallization. In-house data collection is sped up with commercial imaging plates. Now, one can find and refine the cryo-conditions, bring crystal trays to the synchrotron, and freeze directly in the cryo-stream. By and large, MAD phasing is still impractical for the masses and crystallographic software is still a bit difficult to use. The total time to complete the project from protein sample to a fully refined molecule is around 1 to 2 years.

1995 to present time

Now, in-house cloning, expression, purification, crystallization, and incorporation of selenomethionine (SeMet) become the norm. Crystals are often frozen in-house, screened, stored frozen, and then shipped to the synchrotrons for data collection. By treating MAD as a special case of MIR, phasing becomes easy for the masses (Ramakrishnan and Biou [1] as implemented in MLPHARE, CNS/X-PLOR, SOLVE, SHARP, and XtalView). The structure can even be solved at the beamline after using direct methods techniques to find the selenium substructure. From the start of data collection to electron density mapping the time required can be as short as 9 hours. The total time to complete the project from cloning the protein to a fully refined molecule is around 6 months (e.g. see Peat et al. [2]).

1998-2000

Several protein structure initiatives begin, starting first with small pilot projects. In the first protein structure initiatives, thermophiles are targeted because their proteins are more stable, easier to express and purify from bacterial systems, and possibly easier to crystallize. For the first pass, a small number of proteins are cloned and expressed; proteins that crystallize easily are solved, blindly filling out the structural database. Sets of proteins are specifically targeted. One group may try to flesh out the library of protein folds; another may target proteins that cause disease in humans. During this period, we hope to see the structures solved of around 10% of the estimated 2,500 proteins in the different 'proteome' pilot projects. The total time to complete a structure, going from protein sample to a fully refined molecule, will be, optimistically, around 1 month.

2000-2010

There will be a massive deployment of collaborative efforts targeting the 3-D structures of proteins in the human genome-crystallography on a scale similar to the various ongoing genome sequencing projects. Strong government backing by NIH and DOE will develop centers for large-scale expression, purification and crystallization. Synchrotron beamlines will operate round-the-clock shifts of staff scientists. They will receive mailed frozen crystals and collect MAD data. A 1000 MHz DEC-ALPHA or a 800 MHz Intel-Merced Linux computer will do the data processing and phasing entirely at the beamline. A 3x3 CCD with 0.5 second readout and a 300x300 mm active area will replace the 2x2 CCD. A four wavelength SeMet MAD experiment will be finished in a few hours. Data processing will be done on the fly while data is being collected. With the tremendous processing power and storage available, data will be collected in fine slicing mode to obtain accurate 3-D profiles of the reflections. Phasing will be done by black-box programs, like the current SOLVE [3], at the beamline as soon as data collection finishes. The beamline scientists will examine the quality of the maps and transfer the structure factors, phases and maps back to the requesting scientist. At home, using a program that automatically traces electron density, as well as refining the model, a scientist will be able to finish the structure in 3 days. The coordinates, original structure factors, and phases will be checked automatically and deposited at the Protein Data Bank. Ten days after data collection started, the coordinates will be released to the public. Home X-ray sources will be used mostly to screen crystals and find cryo conditions. Small data sets will be collected to estimate the mosaicity and overall quality of crystals that will be catalogued and mailed to the beamline.

2010-2020

Teraflops desktop computers and some clever engineering will allow a robotic arm inside the beamline hutch to open a dewar holding an array of frozen crystals. With cryo-tongs, the robotic arm will fetch a frozen crystal and mount and optically align it on the goniometer. A MAD experiment will begin automatically, and a few hours later, the automatically traced and refined molecule will be sent to the molecular biologist.

At this stage, we will have in place a system philosophically similar to today's small molecule service labs.

2020 to 2050

By the middle of the century all the proteins of the human genome will be solved by a combination of molecular replacement, ab initio phasing and MAD phasing. The rest of the soluble proteins that fail to crystallize will be solved by NMR techniques; electron diffraction techniques will be used to solve 2-D crystals of membrane proteins, as well as initial phasing of very large structures [4]. Diffraction labs will work on molecular modeling. Molecular biologists, chemists, materials scientists and modelers will work together doing molecular design and engineering, solving mutant structures, fitting new pharmacophores to recently solved structures, and solving those difficult structures that will always exist.

So, what's driving these advances? A series of developments from this field and others (CCD detectors, synchrotrons, cryo-crystallography, MAD phasing, SeMet incorporation, improved computers and programs, and the sequencing projects) have all come together synergistically in such a way that rapid structure determination is becoming a reality. To achieve the speed that we would all like to see, we suggest a few steps to streamline the process:

First, at home:

1) Set up a local protein expression facility to do the cloning, expression, and purification of large amounts of SeMet Proteins. This system is already operating at a few universities, including UCLA, where it is provides the backbone for the protein structure initiative [5].

2) Macromolecular crystallization is still the bottleneck. Typically, researchers use a relatively small set of commercially available conditions for each protein. The availability of sequence information from many different organisms means that researchers can now screen the same protein from several organisms, considerably increasing the chances of crystallization. Another option is systematically screening mutants of the protein. We'd like to see the use of crystallization robots with densely packed micro array crystallization plates capable of setting up 10,000 different conditions per day. Conditions could be selected by incomplete factorial or/and combinatorial techniques, also varying the protein species as a parameter. Of course, we'd need a robotic CCD camera and a pattern-recognizing computer to detect crystals and catalogue them at a very rapid rate. With superior cryo systems and powerful beamlines, even microcrystals could be sufficient.

3) Further improvement of in-house x-ray facilities are needed. For example, confocal multilayer optics will give smaller and more collimated beams than present double bended mirrors. Large anode generators such as the Rigaku FRD high flux generator will produce a beam size of 150 microns with a load of 5 kW. As CCD detectors become cheaper to produce, they will replace image plate detectors. CCD detectors are ideal systems for the rapid screening of crystals. These improved facilities can also be used for data collection where phasing is not an issue, e.g. for crystals of mutant proteins or for crystals soaked with ligands.

The home source must become a screening facility with the ultimate goal of sending the best characterized frozen crystals to the beamlines. Specifically, crystals sent to the synchrotron should be well characterized for diffraction resolution and mosaicity (including anisotropy in these parameters). Bringing trays with crystals and finding cryo conditions at the synchrotron is an inefficient use of beamline time. A wide range of freezing options can be screened at home, including trials of various cryo-protectants (including adding the cryoprotectant to the crystallization condition), various transfer techniques, and the small molecule workhorse paratone oil [6]. Crystal annealing [7] may help as well.

4) Finding heavy atoms in-house is time consuming and impractical, and should be used as the last resort. From day one in the project, the scientists should plan to incorporate heavy anomalous-scattering atoms, like SeMet, into the protein (at least one SeMet for every 50 to 100 amino acids). In cases where SeMet is not practical, the expression facility could engineer cysteines for a potential Hg MAD experiment. In the case of DNA binding proteins the local DNA Synthesizer can synthesize nucleic acids derivatized with iodine, bromine or reactive thiol groups. Direct incorporation of heavy atoms via peptide synthesis may become an option for relatively small proteins.

5) In-house training of students and postdocs to use the whole range of modern programs from automatic phase solution (SOLVE, SnB, SHELXS, CCP4) to semi-automatic molecule tracing and refinement (0, XtalView, WARP, REFMAC, CNS, SHELXL) and structure validation (VERIFY3D, ERRAT, PROCHECK, WHAT-CHECK). We suspect students of the future won't be 'crystallographers', but users of crystallographic software and techniques integrated with other biophysical techniques.

6) With the beginning of an avalanche of macromolecular structures, we need to insure the integrity and accuracy of the database entries. The crystallographic community should agree on what standards are required to consider a structure finished. For a given resolution, there should be standards for acceptable values of R_{cryst}, R_{free}, bond length and angle deviations from ideality, overall diffraction data completeness and completeness in the last shell. The editors of journals must force the authors to deposit and release the atomic coordinates, diffraction data and experimental phases before a structure can be published.

And, of course, at the synchrotrons:

1) Beamlines must streamline their operations too. Ideally, each beamline should have crystallographers present around the clock to mount the frozen crystals and collect the MAD data. A tunable beamline with a CCD detector should be able to collect 3 to 5 MAD data sets per day. Since all of the crystal screening is done at home, beamlines could operate in a “remote mode” following the Advanced Light Source (Berkeley) model [8]. This so-called 'Fed-Ex crystallography'' will allow all research groups to ship pre-tested and screened frozen crystals to the beamlines. The crystallographer at the beamline will collect the data following the researcher's instructions. The beamline crystallographer can either send raw images or the completed phasing experiment, with structure factors, phases and a map. In fact, all communications and even some data display could be done in real-time over the internet.

2) Computer power must be increased at least tenfold at the beamline to enable realtime processing of data, phasing, and visualization of electron density maps. Expensive computers with commercial operating systems can be replaced with inexpensive Intel Pentium-II personal computers or DEC-Alphas running Linux.

So the burning question now becomes: If proteins are expressed by professionals at the “Protein Expression Facility', crystals conditions are determined by a robot, and the MAD experiment is performed in 'remote mode” by a professional crystallographer at the beamline, then what is the Molecular Biologist/Crystallographer supposed to do now? Of course, our ultimate goal has always been to find the function of these proteins. The difficulties of structure determination have made many of us mistake the technique for the goal itself. We hope that advances in structure determination will free us to return to trying to understand biology.

Of course, with the advances we've guessed at on the horizon, the face of crystallography will change as well. We'll see increases in the structural databases and all the benefits that come with that, like modeling of structures and mutations, and increased understanding of protein and RNA folding rules. More and more, macromolecular crystallography will break down into two realms. Routine structure determination will expand from today's mutagenesis and ligand binding studies to include projects such as de novo MAD

structure determinations. These techniques will be as accessible to biochemists as any other biophysical method. But we also see crystallographers specializing in attacking non-standard problems (hemihedral twinning, severe anisotropy) or difficult, 'tour-de-force' structures. Massive structures such as transcriptional machinery or the translational apparatus of the ribosome will be solved through synchrotron radiation, clever phasing techniques, and the continued tradition of doggedly ingenious crystallographers.

References

[1] Ramakrishnan, V. and Biou, V. (1997) Treatment of MAD Data as a Special case of MIR. *Methods in Enzymology* **276**: Macromolecular Crystallography, Part A. Charles W. Carter, Jr., and Robert M. Sweet, eds. (http://snowbird. med.utah.edu/~ramak/madms)

[2] Peat, T. S., Newman, J., Waldo, G. S., Berendzen, J. and Terwilliger, T. C. (1998) Structure of translation initiation factor 5A from *Pyrobaculum aerophilum* at 1.75 Å resolution. *Structure* **6**: 1207-1214.

[3] SOLVE: (http://www.solve.lanl.gov)

[4] Ban, N., Freeborn, B., Nissen, P., Penczek, P., Grassucci, R. A., Sweet, R., Frank, J., Moore, P. B. and Steitz, T. A. (1998) A 9 Å resolution X-ray crystallographic map of the large ribosomal subunit. *Cell* **93**: 1105-1115.

[5] UCLA Protein Expression Facility. (http://www.doe-mbi.ucla.edu/People/Perry/)

[6] Rodgers, D.W. (1994) Cryocrystallography. *Structure* **2**: 1135-1139.

[7] Harp, J. M., Timm, D. E. and Bunick, G. J. (1998) Macromolecular crystal annealing: Overcoming increased mosaicity associated with cryocrystallography. *Acta Cryst.* **D54**: 622-628.

[8] Thomas Earnest, Advanced Light Source, Berkeley, Personal Communication.

Kenneth Goodwill Edward Marcotte Duilio Cascio