TITLE: Structural Biology Computing: Lessons For The Biomedical Research Sciences

Authors: Andrew Morin, Piotr Sliz
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston MA 02115

The field of structural biology, whose aim is to elucidate the molecular and atomic structures of biological macromolecules, has long been at the forefront of biomedical sciences in adopting and developing computational research methods. Operating at the intersection between biophysics, biochemistry and molecular biology, structural biology's growth into a foundational framework on which many concepts and findings of molecular biology are interpreted[1] has depended largely on parallel advancements in computational tools and techniques. Without these computing advances, modern structural biology would likely have remained an exclusive pursuit practiced by few, and not become the widely practiced, foundational field it is today. As other areas of biomedical research increasingly embrace research computing techniques, the successes, failures and lessons of structural biology computing can serve as a useful guide to progress in other biomedically related research fields.

Evolution from restrictive to accessible

Since their founding in the first half of the 20th century, the three major complementary experimental techniques of structural biology, X-ray crystallography, nuclear magnetic resonance (NMR) and electron microscopy (EM), have each experienced profound transformations, not only in capabilities and capacity, but also in accessibility. While it is expected that scientific techniques improve and become more widely utilized as they develop and mature, the magnitude and scope of changes experienced by structural biology techniques has led to a fundamental change not only in the 'how' of structural biology research, but also the 'who'.

Improvement in experimental hardware – synchrotron X-ray sources, high-field NMRs and high-voltage EMs – deserve much credit for increased capability and productivity. However, the most profound effects for democratizing and opening structural biology research to a wider, less specialized user base are attributable largely to advances in structural biology computing tools and techniques. Perhaps no better example of this transformation from restrictive to accessible can be found than X-ray crystallography.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1002/bip.22343

Once a highly technical, abstruse and time-consuming process, X-ray crystallography is now an accessible and practical experimental technique routinely performed by scientists from disparate backgrounds and training.
In the 1930s, 40s and 50s, the pioneers of X-ray crystallography solved protein structures by recording X-ray diffraction patterns on sheets of photographic film, performed difficult Fourier and inverse-space calculations by hand[2] or using early rudimentary computers, traced electron densities on stacks of acetate overlays and visualized models using handmade sculptures of wood and wire. The collecting and processing of diffraction data was technically demanding, difficult and slow.
Today, X-ray diffraction is collected digitally by advanced 'direct' detectors[3] and entire computational structure determination workflows, from data collection to processing, phasing, model building, refinement, validation and visualization, can be performed in a few hours on a standard laptop computer by scientists with little formal structural biology training.

_____BEGIN_SIDEBAR_____
Early X-ray crystallographers referred to attaining the 3D coordinates of a protein as "solving" the structure, reflecting their perception of the process as a kind of intricate scientific puzzle of many parts. Today, the preferred terminology is "structure determination" denoting the more prescribed and rigorous methods typically employed.
_____END_SIDEBAR_____

Similarly, NMR has become both more capable and accessible due largely to developments in computing. Biomolecular NMR spectroscopy can be used to determine the 3D structure, dynamics and interactions of proteins, nucleic acids and other macromolecules. Structure determination in NMR involves multiple experimental data collection, processing and analysis steps and depends on computing at every stage, including converting raw data into interpretable spectra, assigning chemical shifts, computing dipolar couplings and calculating structure from restraints.
Possibly due to the greater variety of experiment types and endpoints, or the more constrained availability of high-field spectrometers, biomolecular NMR has not yet achieved the widespread level of adoption and ease-of-use advancements of X-ray crystallography. Significant strides have been made however, in the computational automation or semi-automation of tedious and time-consuming steps such as resonance assignment and calculation of restraints[4-6]. This has resulted in greater efficiency and productivity for expert users as well as lowered barriers to entry for users with more general biology backgrounds.
EM has likewise experienced significant progresses enabled by parallel advancements in computing. Structural biologists have utilized EM from its beginnings in the 1930s. Only since the development of cryoEM techniques in the 1980s[7] however, has high-resolution imaging of biomacromolecules become possible. State-of-the-art recent advances have pushed cryoEM structures to resolutions approaching that of X-ray crystallography[8,9].

These EM techniques require large-scale computing resources for the storage, processing and reconstruction of millions of individual particle images into a high-resolution 3D model. Though recently developed and still rapidly evolving, advances in cryoEM software, processing and automation methods have greatly expanded the practical reach of the technique. Software for creating image processing workflows and exploiting parallel processing resources have also reduced the computational expertise required of those making use of EM to attain structure information.

Advancement through collaboration and competition

Much of the progress in capability, productivity and accessibility in structural biology computing can be ascribed to advancements in software. Software is the mean by which scientists harness the power of computing. The open exchange and availability of structural biology software through collaborative efforts enhances productive competition. Facilitated competition through collaboration has been a major driving force for innovation and improvement in structural biology computing.

In the 1950's and 60's, as mainframe computers became available at select institutions, researchers began creating programs to accelerate time and labor-intensive steps in structure calculation. Researchers ability to share programs between groups or institutions, however, was impeded by incompatible computer architectures; ad hoc software created for ad hoc hardware[10]. As access to computing resources spread and computing architectures began to standardize in the late 1960's and 70's, the ability to share and directly compare programs employing different implementations of underlying computational methods and algorithms stimulated productive competition amongst structural biology program developers.

Recognizing the merits of sharing computational tools and methods, software collaborations formed to foster scientific exchange, prevent unproductive redundancies and aid discovery and dissemination of structural biology programs. The first among these was the Collaborative Computing Project 4 (CCP4) created in 1979[11]. The CCP4 software suite is a collection of complementary and competing programs created by structural biology researchers for various stages of the X-ray diffraction structure determination process. CCP4 has since grown to incorporate more than 250 individual programs. Creation of a graphical user interface (GUI) called CCP4i[12] linking many of the programs along with recent efforts to prioritize validation and automation have made CCP4 programs accessible to a wide range of users without diminishing utility for expert crystallographers.

The Collaborative Computing Project for NMR (CCPN)[13] initiated in 1999 has taken a similar approach to expanding access and promoting competition and cooperation among programs for NMR research.

A contrasting example of an X-ray crystallography software collaboration with a somewhat different design philosophy is the PHENIX software suite[14]. While collaborative, and incorporating programs from multiple sources, PHENIX is an attempt to mitigate some of the drawbacks to all inclusive program collections like

CCP4. Utilizing consistent programming languages and conventions, the PHENIX design philosophy is to implement the leading structure determination method at each step of a closely integrated yet flexible structure determination workflow. PHENIX's emphasis on automation of standard structural workflows permits novice users to more easily process and validate X-ray structure data, while also allowing efficient inclusion of advanced features and methods.

The merits and drawback of the two approaches to structure determination software suites, the inclusiveness of CCP4 or the tight integration of PHENIX, is a matter open for debate and crystallographers commonly utilize both program suites. Nonetheless, productive competition both within and between software suites benefits structural biology users and program developers, with advances by one often spurring progress by the other.

Still another approach to software collaboration is exemplified by the SBGrid Consortium[15]. SBGrid is an example of a community-based software collaborative providing comprehensive support for all types of structural biology computing. The SBGrid software collection contains CCP4, PHENIX, CCPN, plus 270 other software suites and individual programs in areas covering X-ray crystallography, NMR, EM, bioinformatics, cheminformatics, molecular dynamics (MD), structure prediction and others. In addition to assembling a comprehensive set of structural biology software, SBGrid supports use of the software by compiling, configuring, installing and updating the entire collection automatically on computers in member laboratories without need for end-user action. Thus, members of SBGrid have convenient and unrestricted access to multiple competing applications for each aspect of a structural biology workflow, and can easily evaluate and compare computational tools and methods.

A specific example of the beneficial effects of competition between different programs implementing the same method can be seen in the evolution of the molecular replacement (MR) technique in X-ray structure determination.

Solving the "phase problem" has been a difficulty for crystallographers since the origins of X-ray diffraction. Historically, isomorphous replacement[16] by introduction of heavy atoms into protein crystals was the only means of determining the phase of diffracted X-rays, and thereby convert diffraction data into real-space 3D protein structures. Isomorphous replacement or similar anomalous dispersion methods[17] remain the preferred means of phasing data from a protein with novel structure. However, if a model of the protein under investigation can be obtained from either a close homolog or by predictive protein modeling, MR methods can then back calculate diffraction patterns and use this information to position the model. If this step is successful MR then uses the model information to phase the experimental data. Rossmann's initial formulations for molecular replacement[18] have been utilized since the early 1960's to improve experimentally determined phases and for phase determination. However, it was not until the late 1980's, coinciding with improvements in computing power and availability, that molecular replacement came into widespread use.

The first widely distributed software suite[19] for determining protein structure by MR called MERLOT[20] released in the late 1980s was followed soon thereafter by increasingly capable programs like AMORE[21], Molrep[22] and others. To date, over a

dozen different MR programs have been created for X-ray crystallographic structure solution, each of which take different approaches to implementing the basic method described by Rossmann[23], with the newer programs frequently eclipsing older in both efficacy and ease of use. MR has also been adapted to make use of the potential offered by large-scale computing. Wide search molecular replacement (WSMR) uses massively-parallel computing resources to perform MR of experimental data against a database of all known protein folds, families and superfamilies[24]. WSMR holds the potential to phase structures independent of sequence information or known structure homology.

The potential to skip the difficult and time-consuming experimental steps of isomorphous replacement or anomalous dispersion, and rapid improvements in speed, accuracy and ease of use of MR methods thanks in part to robust competition between programs, has led to the adoption of MR as a standard structure determination technique. The current most popular program for MR, Phaser[25], has been incorporated into both CCP4 and the PHENIX program suites.

Standardization promotes competition and access

An oft overlooked but key component to maximizing the benefits of shared resources, competition and collaboration are the setting of standards to address recognized community needs. Top down attempts to develop and enforce standards often fall short or suffer from unintended effects, while individual efforts are frequently too narrow to achieve widespread acceptance. Conversely, initiatives based around representative communities of stakeholders have generally proven more successful.

In addition to offering a suite of NMR software, CCPN has made strides in establishing common NMR data standards to promote interoperability between different programs used in the various stages of data collection and processing[26].

Perhaps the best-known and most widely utilized computational resource in structural biology is the worldwide Protein Data Bank (PDB). Today deposition of structure data to the PDB is a prerequisite for publication[27] and the PDB offers free online access to a searchable database of over 90,000 3D structures of proteins, nucleic acids and their complexes. Its central role as worldwide repository for biomacromolecular structures [28] has allowed the PDB to facilitate extensive standardization in the deposition of structure data, including reporting requirements for processed and raw experimental data, methods and protocols, data formatting and other information useful in both validating and reproducing structure determination and in promoting compatibility and interoperability. The standard '.pdb' file format for describing the 3D structure and other metadata of macromolecules is widely used not only in the three main techniques of structural biology, X-ray crystallography, NMR and EM, but also in fields as diverse as chemistry, bioinfomatics, molecular dynamics, protein structure prediction and others. Near-universal adoption of the .pdb file format among molecular visualization software has also benefitted researchers ability to utilize and compare multiple applications. Currently, the PDB is coordinating the transition to a more

modern file format (PDBx/mmCIF) designed to accommodate recent advances in computing and experimental data collection.
The PDB has also been integral to efforts to develop standard metrics for quality assessment of structures determined by X-ray diffraction methods. These metrics, and methods for calculating them, have been incorporated into programs and software used in structure determination, as well as an additional validation step in the PDB deposition process.

Scientist-created software and dissemination

An important lesson from the history of structural biology computing is that aiding and encouraging scientists to create computational tools and techniques is vital to progress. Nearly all of the computational resources discussed above were created by structural biologists, for structural biologists, and the most widely used and influential tools, programs, collaborations and communities were built by practicing scientists.
Due largely to historical origins steeped in mathematics and physics, structural biology continues to profit from a greater representation of researchers skilled in programming than most other biomedical research fields. Policies and practices that promote programming education among practitioners would seem likely to pay similar dividends in the development and exploitation of computing resources for other scientific fields.
Equally important as promoting tool creation is the effective dissemination and support of computational tools and techniques created by scientists. Programs and software produced in the course of research are typically described in the primary scientific literature, where researchers wishing to share their programs often provide links to laboratory websites for downloading. Discovery of a particular program is dependent on potential users reading the primary literature, word of mouth, presentations at scientific conferences, etc. This is generally not an efficient way to promote the spread and adoption of a given program.
Software collaboratives such as CCP4 provide a much more effective and efficient means of dissemination software by creating a ‘one stop shop’ for research software. Precompiling the components of the software collection for various popular computer operating systems significantly eases the burden on end-users of compiling and installing each individual program, and coordinating component programs under a single software license agreement saves end-users and their institution’s legal counsels time and effort[29].
The SBGrid Consortium goes even further in supporting the dissemination and use of structural biology research computing. In addition to assembling, curating and precompiling a comprehensive collection of software for all major structural biology computing techniques, SBGrid also automatically installs and regularly updates the entire software collection directly on computers in member laboratories via the internet, without need for user action. In this fashion SBGrid relieves researchers from the burdens of software self-support, removes impediments to software

dissemination, lowers barriers and expands access to structural biology computing resources and promotes program discovery.

Large-scale computing: the enabling impediment

Scientist created programs and computational tools are a primary enabling factor in structural biology computing, however data and software go hand in hand. Structural biology was the original 'big data' biomedical science, commonly generating megabytes of data long before magnetic digital storage able to contain it became practical. Yet as structural biology techniques continue to progress and generate ever-greater amounts of data, storing, transmitting and processing this data remains a significant difficulty in structural biology computing.

A typical crystallographic data collection at a regional X-ray synchrotron facility 10 years ago may have generated hundreds of megabytes. Today, a similar experiment may generate a few tens of gigabytes of raw data. This increase in data generation is similar in magnitude and rate for many biomolecular NMR experiments. Though not extreme by today's 'big-data' standards, these datasets can nonetheless pose difficulties in transport, long-term archival storage and sharing for research groups engaging in multiple data collections per month. Prototype solutions making use of internet and national data infrastructure have been demonstrated that provide efficient and secure transmission of data sets from site to site, long-term, high-reliability archiving and data sharing[30]. However, no standard for large-scale data handing has yet gained widespread adoption, and ad hoc solutions are the norm.

Though the scale of raw experimental data from X-ray and NMR techniques can sometimes pose problems, cryoEM is currently the reigning champ of data generation in structural biology. High-resolution 3D reconstructions of proteins using cryoEM often require the collection of millions of individual particle images. Raw cryoEM images are typically produced at the rate of tens of gigabytes per experiment, per day. Large EM laboratories can easily generate dozens of terabytes of data per month. Collected data must be available for computational processing to produce 3D structures. Large-scale, high-reliability storage solutions adopted from the computing and technology industries exist, but are difficult and expensive to deploy and maintain, and as such can present a significant barrier to entry.

For some structural biology techniques, Moore's law decreases in cost and increases in storage capacity have kept pace or exceed the growth in experimental data generation. For others, like cryoEM, practical and economical long-term data solutions are still being sought.

Like the storage and transmission of experimental data, the raw computational power required to perform many new and emerging structural biology techniques is equally formidable in scale and demanding in specialized technical knowledge.

A single near-atomic resolution cryoEM particle reconstruction can require tens of thousands of processor hours to complete[9]. Performing these calculation on a desktop computer workstation would require many years. High performance computing (HPC) clusters, distributed and grid computing approaches are currently the only feasible way to obtain required computing power. The knowledge and skills

to effectively implement and exploit these large-scale computing approaches is rare however, even among computationally proficient structural biologists, thereby rendering these new computational techniques inaccessible to most would-be users. One might expect that, as with other structural biology techniques, these new large-scale computing techniques will also evolve over time from difficult and restrictive to accessible and practical, and indeed, this is already beginning to occur.
Science computing portals simplify and expand access to large-scale computing resources. Concise web-based interfaces mask the complex process of preparing and distributing user submitted data, launching computations on remote HPC infrastructures, monitoring, collecting, assembling, analyzing and delivering results. The WeNMR[31] project provides access to large-scale computational resources necessary for NMR and SAXS data analysis and structural modeling. Similarly, ROSIE[32] facilitates use of various ROSETTA protein structure prediction and design capabilities. Together with the Open Science Grid (OSG) [33,34], SBGrid maintains portals for refinement of low-resolution electron density data[35] and wide-search molecular replacement[24] accessible to the entire structural biology community. Science portals expand access to computing resources that might otherwise be out of reach to most structural biology researchers. It is more than slightly ironic however, that in this era of expanded access and practice of structural biology methods and techniques enabled largely by advances in computing, the latest advances in computational methods present as both chief impediment, and most likely path, to progress.

Why look to structural biology?

Structural biology is far from the only experimental biomedical research field to have adopted computational tools and techniques. Genetics routinely generates larger datasets. Proteomics relies heavily on computing methods to process raw data into interpretable results. Yet of all the biomedical disciplines to have embraced computing, structural biology is arguably the most diverse and heterogeneous. Though its origins and much ongoing research and development lay in its biophysical roots, structural biology today is comprised of science and scientists from all domains of biomedicine. Structural biology computing has, from necessity, adapted to incorporate a broader range and variety or experimental data, equipment, workflows, research goals, backgrounds and training than other fields. Experimental techniques commonly employed in structural biology span disciplines from microbiology to particle physics, and experimental equipment range in size form benchtop appliances to kilometer-wide synchrotron facilities. Structural biology computing has likewise developed over time to serve the needs of this uniquely heterogeneous scientific environment, and adapted to meet the needs of its diverse practitioners. Because of this, the lessons of structural biology computing are likely to hold unusual relevance for other diverse areas of experimental biomedical science as they too increasingly embrace computing.

REFERENCES

(1) Campbell, I. D. *Nat. Rev. Mol. Cell Biol.* **2002**, *3*, 377–381.
(2) Harvard University - Department of History of Science: Beevers-Lipsom Strips http://dssmhi1.fas.harvard.edu/emuseumdev/code/emuseum.asp?emu_action=searchrequest&newsearch=1&moduleid=1&profile=objects&currentrecord=1&searchdesc=Beevers-Lipson%20strips%20in%20oak%20case&style=single&rawsearch=id/,/is/,/1172/,/false/,/true (accessed Jun 14, 2013).
(3) Gruner, S. M. *Phys. Today* **2012**, *65*, 29–34.
(4) Güntert, P. In *link.springer.com*; Humana Press: New Jersey, 2004; Vol. 278, pp. 353–378.
(5) Guerry, P.; Herrmann, T. *Methods in molecular biology (Clifton, N.J.)* **2012**, *831*, 429–451.
(6) Shen, Y.; Vernon, R.; Baker, D.; Bax, A. *Journal of biomolecular NMR* **2008**, *43*, 63–78.
(7) DUBOCHET, J. *Journal of Microscopy* **2011**, *245*, 221–224.
(8) Bai, X.-C.; Fernandez, I. S.; McMullan, G.; Scheres, S. H.; Kühlbrandt, W. *eLife Sciences* **2013**, *2*.
(9) Grigorieff, N.; Harrison, S. C. *Current opinion in structural biology* **2011**, *21*, 265–273.
(10) Meyer, E. F. *Protein science : a publication of the Protein Society* **1997**, *6*, 1591.
(11) Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. *Acta crystallographica. Section D, Biological crystallography* **2011**, *67*, 235–242.
(12) Potterton, E.; Briggs, P.; Turkenburg, M.; Dodson, E. *Acta crystallographica. Section D, Biological crystallography* **2003**, *59*, 1131–1137.
(13) Vranken, W. F.; Boucher, W.; Stevens, T. J.; Fogh, R. H.; Pajon, A.; Llinas, M.; Ulrich, E. L.; Markley, J. L.; Ionides, J.; Laue, E. D. *Proteins* **2005**, *59*, 687–696.
(14) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. *Acta crystallographica. Section D, Biological crystallography* **2010**, *66*, 213–221.
(15) SBGrid - Software Consortium for Structural Biology http://www.sbgrid.org/ (accessed Aug 13, 2012).
(16) Green, D. W.; Ingram, V. M.; Perutz, M. F. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **1954**, *225*, 287–307.
(17) Hendrickson, W. A.; Smith, J. L.; Phizackerley, R. P.; Merritt, E. A. *Proteins* **1988**, *4*, 77–88.
(18) Rossmann, M. G.; Blow, D. M. *Acta Cryst* **1962**, *15*, 24–31.

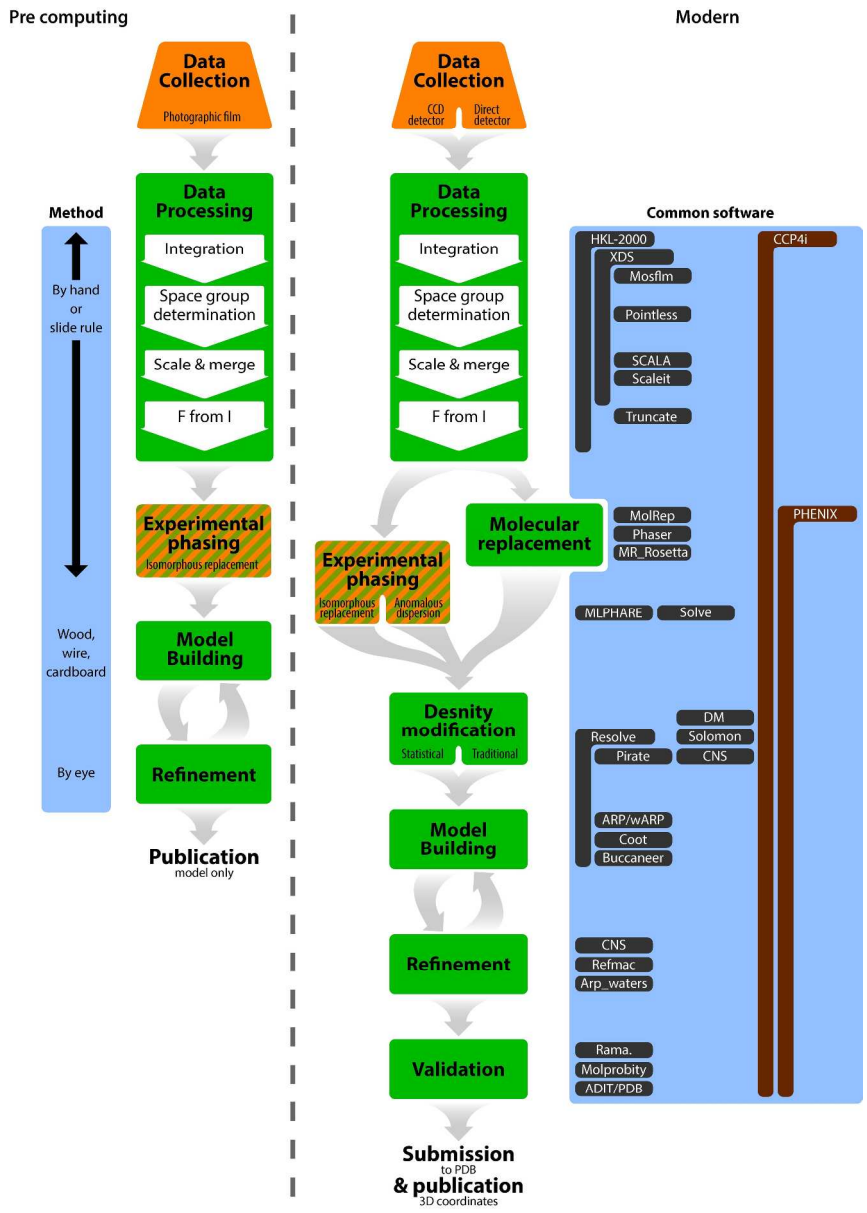
(19) Molecular Replacement Guide http://xray0.princeton.edu/~phil/Facility/Guides/MolecularReplacement.html (accessed Jun 14, 2013).
(20) Fitzgerald, P. M. D. *Journal of Applied Crystallography* **1988**, *21*, 273–278.
(21) Navaza, J. *Acta crystallographica. Section A, Foundations of crystallography* **1994**, *50*, 157–163.
(22) Vagin, A.; Teplyakov, A. *Journal of Applied Crystallography* **1997**, *30*, 1022–1025.
(23) Rossmann, M. G. *The molecular replacement method*; Gordon and Breach, 1972.
(24) Stokes-Rees, I.; Sliz, P. *Proceedings of the National Academy of Sciences* **2010**, *107*, 21476–21481.
(25) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. *Journal of Applied Crystallography* **2007**, *40*, 658–674.
(26) Fogh, R.; Ionides, J.; Ulrich, E.; Boucher, W.; Vranken, W.; Linge, J. P.; Habeck, M.; Rieping, W.; Bhat, T. N.; Westbrook, J.; Henrick, K.; Gilliland, G.; Berman, H.; Thornton, J.; Nilges, M.; Markley, J.; Laue, E. *Nature structural biology* **2002**, *9*, 416–418.
(27) Morin, A.; Urban, J.; Adams, P. D.; Foster, I.; Sali, A.; Baker, D.; Sliz, P. *Science (New York, N.Y.)* **2012**, *336*, 159–160.
(28) Berman, H.; Henrick, K.; Nakamura, H. *Nature structural biology* **2003**, *10*, 980–980.
(29) Morin, A.; Urban, J.; Sliz, P. *PLoS computational biology* **2012**, *8*, e1002598.
(30) Stokes-Rees, I.; Levesque, I.; Murphy, F. V.; Yang, W.; Deacon, A.; Sliz, P. *J Synchrotron Radiat* **2012**, *19*, 462–467.
(31) Wassenaar, T. A.; Dijk, M.; Loureiro-Ferreira, N.; Schot, G.; Vries, S. J.; Schmitz, C.; Zwan, J.; Boelens, R.; Giachetti, A.; Ferella, L.; Rosato, A.; Bertini, I.; Herrmann, T.; Jonker, H. R. A.; Bagaria, A.; Jaravine, V.; Güntert, P.; Schwalbe, H.; Vranken, W. F.; Doreleijers, J. F.; Vriend, G.; Vuister, G. W.; Franke, D.; Kikhney, A.; Svergun, D. I.; Fogh, R. H.; Ionides, J.; Laue, E. D.; Spronk, C.; Jurkša, S.; Verlato, M.; Badoer, S.; Dal Pra, S.; Mazzucato, M.; Frizziero, E.; Bonvin, A. M. J. J. *J Grid Computing* **2012**, *10*, 743–767.
(32) ROSIE: The Rosetta Online Server that Includes Everyone http://rosie.graylab.jhu.edu/ (accessed May 6, 2013).
(33) Stokes-Rees, I.; O'Donovan, D.; Doherty, P.; Porter-Mahoney, M.; Sliz, P. IEEE; pp. 1–8.
(34) Pordes, R.; Petravick, D.; Kramer, B.; Olson, D.; Livny, M.; Roy, A.; Avery, P.; Blackburn, K.; Wenaus, T.; Würthwein, F.; Foster, I.; Gardner, R.; Wilde, M.; Blatecky, A.; McGee, J.; Quick, R. *J. Phys.: Conf. Ser.* **2007**, *78*, 012057.
(35) O'Donovan, D. J.; Stokes-Rees, I.; Nam, Y.; Blacklow, S. C.; Schröder, G. F.; Brunger, A. T.; Sliz, P. *Acta crystallographica. Section D, Biological crystallography* **2012**, *68*, 261–267.
(36) Otwinowski, Z.; Minor, W. *Methods in enzymology* **1997**.
(37) Kabsch, W. *Acta crystallographica. Section D, Biological crystallography* **2010**, *66*, 125–132.
(38) Battye, T. G. G.; Kontogiannis, L.; Johnson, O.; Powell, H. R.; Leslie, A. G. W.

Acta crystallographica. Section D, Biological crystallography **2011**, *67*, 271–281.
(39) Evans, P. *Acta crystallographica. Section D, Biological crystallography* **2005**, *62*, 72–82.
(40) Diederichs, K.; Karplus, P. A. *Nature structural biology* **1997**, *4*, 269–275.
(41) Howell, L.; Smith, D. *J.Appl. Cryst.* **1992**, *25*, 81–86.
(42) Wilson, K. S.; French, G. S. *Acta crystallographica. Section A, Foundations of crystallography* **1978**, *34*, 517.
(43) DiMaio, F.; Terwilliger, T. C.; Read, R. J.; Wlodawer, A.; Oberdorfer, G.; Wagner, U.; Valkov, E.; Alon, A.; Fass, D.; Axelrod, H. L.; Das, D.; Vorobiev, S. M.; Iwaï, H.; Pokkuluri, P. R.; Baker, D. *Nature* **2011**, *473*, 540–543.
(44) Otwinowski, Z. 1991; Vol. 25, p. 26.
(45) Terwilliger, T. C.; Berendzen, J. *Acta crystallographica. Section D, Biological crystallography* **1999**, *55*, 849–861.
(46) Cowtan, K. D.; Main, P. *Acta Crystallographica Section D* **1996**, *52*, 43–48.
(47) Abrahams, J. P.; Leslie, A. G. W. *Acta crystallographica. Section D, Biological crystallography* **1996**, *52*, 30–42.
(48) Cowtan, K. *Acta crystallographica. Section D, Biological crystallography* **2010**, *66*, 470.
(49) Brunger, A. T.; Adams, P. D.; Clore, G. M. *... Crystallography* **1998**.
(50) Perrakis, A.; Morris, R.; Lamzin, V. S. *Nature structural biology* **1999**, *6*, 458–463.
(51) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. *Acta crystallographica. Section D, Biological crystallography* **2010**, *66*, 486–501.
(52) Cowtan, K. *Acta crystallographica. Section D, Biological crystallography* **2006**, *62*, 1002–1011.
(53) Murshudov, G. N.; Skubák, P.; Lebedev, A. A.; Pannu, N. S.; Steiner, R. A.; Nicholls, R. A.; Winn, M. D.; Long, F.; Vagin, A. A. *Acta crystallographica. Section D, Biological crystallography* **2011**, *67*, 355–367.
(54) Lamzin, V. S.; Wilson, K. S. In *Macromolecular Crystallography Part B*; Charles W Carter, R. M. S., Jr, Ed. Methods in Enzymology; Academic Press, 1997; Vol. 277, pp. 269–305.
(55) Hollingsworth, S. A.; Karplus, P. A. *BioMolecular Concepts* **2010**, *1*, 271.
(56) Chen, V. B.; Arendall, W. B., III; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. *Acta crystallographica. Section D, Biological crystallography* **2009**, *66*, 12–21.
(57) ADIT Deposition Tool http://deposit.rcsb.org/adit/ (accessed May 6, 2013).

FIGURE LEGENDS

Figure 1. Comparison of standard historical and modern X-ray crystallography workflows. Left of hashed line: Typical workflow for protein structure determination by X-ray diffraction before the widespread availability of standardized computing resources. Extensive calculations for the data processing

and experimental phasing steps were performed by hand or ad hoc computer program (blue box at left). Model building was physical process often performed with wood or wire and structure refinement was performed by eye. Publication of early protein structures consisted only of artistic renderings of the protein model, and often did not including 3D coordinates. Orange represents experimental steps; green computational steps; striped denotes steps that are both computational and experimental. Right of hashed line: Standard workflow for modern X-ray diffraction structure solution. Molecular replacement techniques, when possible, offer the potential to skip the often difficult and time-consuming experimental phasing step. Blue box at right displays the stages at which commonly utilized individual programs (dark grey lines) are employed. Vertical dark grey lines denote programs useful at multiple stages. Brown lines denote software suites. Submission of 3D structure data is now a prerequisite to publication.
Programs listed include: HKL-2000[36], XDS[37], Mosflm[38], Pointless[39], SCALA[40], Scaleit[41], Truncate[42], Molrep[22], Phaser[25], MR_Rosetta[43], MLPHARE[44], Solve/Resolve[45], DM[46], Solomon[47], Pirate[48], CNS[49], ARP/wAPR[50], Coot[51], Buccaneer[52], Refmac[53], Arp_waters[54], Ramachandran plots[55], Molprobity[56] and ADIT[57].



271x380mm (300 x 300 DPI)