5. High Throughput Protein Crystallography

Bernhard Rupp

Department of Biochemistry & Biophysics, Texas A&M University, College Station, TX 77843-2128, USA

RUNNING TITLE: High Throughput Crystallography.

Permanent Address : Macromolecular Crystallography and Structural Genomics Group, Lawrence Livermore National Laboratory, University of California, Livermore, CA 94551

br@llnl.gov

Phone (925) 209–7429, Fax (801) 880-3982

Table of contents

5.1. Background and rationale

5.1.1. Definition and scope

The chapter 'High Throughput Crystallography' aims to introduce the reader with general interest in proteomics to the methods and techniques of macromolecular crystallography. Special emphasis will be placed on the challenges faced and the progress made in achieving high throughput in this predominantly used method of protein structure determination.

The scope of this chapter is the entire process of a crystallographic structure determination, beginning from the presence of a crystal of suitable size and morphology (but of unknown quality as far as diffraction is concerned) up to initial validation of the refined molecular model. Chapter 5 thus will not include the arguably most crucial part of a structure determination, namely the production of suitable, often engineered, protein targets (Chapter 1), nor will it cover the actual crystal growth experiments (Chapter 3). Included in the discussion will be procedures ranging from crystal harvesting, cryo-protection and derivatization, through data collection and phasing, to the cycles of model building, refinement, and initial validation. These latter steps constitute - or at least should constitute - an integrated part of the process of structure determination. The extensive array of structure validation and analysis based on structural and chemical plausibility and related prior knowledge warrant their own full chapter 10. Figure (1) provides an overview of the most important steps and methods in the course of a crystallographic structure determination.

5.1.2. Choosing crystallography for high throughput structure determination

The motivation to use X-ray crystallography as the primary means of structure determination (out of ~20k PDB structure entries in February 2003 ~17k are crystal structures, ~3k NMR models) lies mainly in the fact that accurate and precise molecular structures, sometimes at near atomic resolution, can be obtained rapidly and reliably *via* X-ray crystallography. Elucidating the precise atomic detail of molecular interactions is essential for drug target structures to be useful as leads in rational drug design (1). Another advantage of crystallographic structure determination is, that no principal difficulty limits the accurate description of very large structures and complexes, as evidenced in the nearly 2 MDa structure of the 50S ribosomal subunit (2), determined at 2.4 Å resolution. As the emphasis in proteomics shifts towards obtaining a comprehensive picture of protein interactions (3) the capability to determine large, multi-unit complex structures will become increasingly important.

The price to be paid for obtaining high quality X-ray structures is that a diffracting protein crystal needs to be produced. Crystal growth in itself can prove quite challenging (Chapter 3) and often can only be achieved after substantial protein engineering efforts (Chapter 2). Comparison with NMR data confirms that the core structure of proteins remains unchanged (4), and enzymes packed in crystals even maintain biological activity, often necessitating the design of inactive enzyme substrate analogues in order to dissect the molecular reaction mechanisms. The fact that protein molecules are periodically packed in a crystal places limitations on the direct observation of processes involving large conformational changes, which would destroy the delicate molecular packing arrangement necessary to form a protein crystal. Molecular transport processes or interactions involving extended conformational rearrangements may require multiple, stepwise 'snapshot' structure determinations in order to understand such inherently dynamic processes. Time resolved crystallography, although exciting for enzymatic reactions involving limited, local structural changes (5), presently plays no role in high throughput structure determination projects due to substantial technical challenges and limitations in applicability.

5.1.3. Why high throughput?

At this point, it may also be appropriate to recall the motivation behind the desire to achieve high throughput in protein crystallography (as discussed in the wider context of Proteomics in Chapter 1). To a large degree, the driving incentive behind structure determination on a 'genome scale' is the elucidation of the structural basis for molecular function, not just for one enzyme, but for whole functional pathways. Such aspects will become even more important as the classic single gene - single disease targets become increasingly sparse (6). The goal of producing lead structures suitable for therapeutic drug design requires that such structures are of high quality, providing enough accurate detail for *in silico* screening or rational compound design (1). Structures obtained for drug design purposes will naturally have much higher criteria for precision compared to structures determined with the goal to populate the protein fold space, an explicit aim of publicly funded structural genomics projects (7). In many cases, structures of multiple orthologs, numerous functional mutants, or multiple potential drug-protein complexes have to be determined. The need for high throughput, without compromise in structure quality, becomes quite evident.

5.1.4. Definition of High Throughput Protein Crystallography

The term 'High Throughput Protein Crystallography' (HTPX) is used rather loosely in the scientific literature, and comprises work ranging from modest scale, academic efforts focusing on affordable automation (8) to truly industrial scale, commercially driven ventures processing hundreds of crystals a week (9-11). The underlying physical and methodical principles, however, are the same for HTPX as for conventional 'Low Throughput' (LT) protein crystallography. Most of the improved methods and techniques developed for high throughput purposes are also applicable and indeed very useful for LT efforts. It appears fair to say that protein crystallography in general has made significant technical progress through increased public and commercial funding devoted to the development of high throughput technologies since the late 1990s.

5.1.5. High throughput vs. low throughput

The most significant difference separating 'true' HTPX and conventional LT crystallography efforts is the amount of automation and process integration implemented. In contrast to computational prowess, full robotic automation nearly always comes at a substantial cost, and here lies probably the most significant rift between academic low- to medium throughput and commercial HTPX ventures. Financial, infrastructural, and production related constraints, for example, are very different in academic and commercial environments, as are efficiency considerations. In a high throughput environment it may be neither necessary nor efficient to pursue every recalcitrant target to completion, while in an academic setting careers may depend on determining one specific structure. A review of the most significant differences between academic and industrial high throughput efforts in terms of philosophy and efficiency considerations is provided elsewhere (8).

5.1.6. Key developments

High throughput in crystallography has become possible through two major interdependent developments: A rapid progress in technology, in particular advances in cryo-, synchrotron, and computational techniques, as well as the influx of substantial public and venture funding. Practically all true high throughput efforts depend on powerful 3rd generation synchrotron X-ray sources (Table 1), largely because of the high brilliance of the X-rays (important for small crystals) and the unique tunability of the wavelength (12). The capability to choose the wavelengths - in contrast to the fixed, characteristic X-ray wavelength defined by the anode material of conventional X-ray sources - is the basis for anomalous phasing techniques that dominate high throughput protein crystallography (13). Given the potentially enormous rewards of structure guided drug development (14), it comes as no surprise that a substantial number of commercial ventures were able to attract funds to develop and implement high throughput techniques, in particular advanced robotic automation. On the other hand, the recent public NIH-NIGMS funding of structural genomics pilot projects (7) provides, for the first time on a reasonable scale, the means for the development of non-proprietary high throughput

structure determination methods, which has benefited not only the funded projects, but practically every structural biology effort.

5.1.7. Literature

Excellent monographs, series, and journal special issues reflecting the progress in crystallographic techniques and their implication for structural genomics over the recent years are available; some of them are listed in the general reference section (5.4.1). Many citations throughout the text refer to other reviews or review-like articles, which contain further specific technical primary references. This chapter emphasizes aspects of particular relevance to automation and HT protein crystallography. The development in this field is extraordinarily rapid, and to obtain a complete and current picture it will be necessary to supplement the information provided in this article with electronic literature and web site searches.

5.2. Methods of high throughput protein crystallography

5.2.1. Overview

Crystallographic protein structure determination centers around a conceptually quite simple diffraction experiment: A cryo-cooled crystal is placed on a goniostat and exposed to intense and collimated X-rays. The goniostat allows the crystal to be rotated in small increments, and for each orientation a diffraction pattern is collected. The images are indexed, integrated and scaled, and unit cell and space group are determined. A reduced and hopefully complete data set, essentially representing a periodically sampled reciprocal space transform of the molecules comprising the crystal, is obtained. Due to the lack of phase information in the diffraction patterns, direct reconstruction of the electron density of the molecules via Fourier transforms is not generally possible for

proteins (known as the 'Phase Problem' in crystallography, Figure 6). In the absence of a suitable known structure model, additional data sets of isomorphous derivative crystals and/or anomalous datasets at suitable wavelengths need to be collected to allow the determination of phases (detailed in section 5.2.4). Overall, at least 2/3 of all HTPX structures are solved by molecular replacement techniques exploiting a homologous structure or model as a source of initial phases, and the *de novo* phasing[1] of the remaining 1/3 relies heavily on anomalous techniques.

The actual phase calculations, electron density reconstruction, model building and structure refinement are conducted *in silico* with computer programs. The procedure generally begins with determination of a heavy (marker) atom substructure, calculation of initial phases, phase improvement by density modification techniques, and model building and refinement. The last 2 steps are generally used in iteration, with improvements until convergence is achieved. Nearly all of the computational methods have been highly automated, and they are currently being integrated into fully automated structure determination packages (Table 2). Although challenges remain, particularly in model building at low resolution, improved computational methods are continuously developed and tend to perform well.

With respect to high throughput requirements, the major steps in a crystallographic structure determination can be grouped as indicated by the different shading in Figure 1:

- The first group includes processes and steps involving manipulation of the fragile 'raw' crystals, such as harvesting, cryo-protection, soaking, and actual mounting of the crystals on the diffractometer (section 5.2.2).
- In the second group, processes from initial evaluation of the now cryo-protected crystals to completion of data collection are included (section 5.2.3).

[1] I follow the notation of '*de novo*' phasing if no previous protein structure model has been used (thus excluding molecular replacement techniques or difference map techniques). *Ab iniito* phasing refers to the use of (direct) methods which derive the phases solely from the intensities of a diffraction data set.

- Group three comprises the entirely computational steps from phasing to the analysis of the final structure model. No more manipulation of crystals is necessary (section 5.2.4).

The technical challenges faced in automated mechanical manipulation of crystals are fundamentally different from the computational requirements later in the structure determination process. Robotic micromanipulation, particularly harvesting, soaking and cryo-cooling of the fragile protein crystals is very expensive to automate and remains a hurdle for full process automation. As crystals are becoming more plentiful, cryo-mounting may eventually develop into a rate limiting step, and demonstrated success of high throughput crystallography at that point may well justify further substantial investment in high throughput robotic crystal harvesting and in-situ diffraction screening techniques.

At the other extreme, the continuous increase in computational power - still roughly doubling every 18 months according to Moore's law (15) - at decreasing cost and the public funding of development efforts for powerful software packages (Table 3) have made automation of the final computational part of structure determination quite successful, although considerable challenges remain in the area of data collection expert systems and automated model building and completion.

5.2.2. Processes involving crystal manipulation

Micromanipulations during harvesting, mounting, derivative soaking, and cryo-protection present serious challenges for full robotic automation and for achieving sustained high throughput. Although crystal harvesting in suitable cryo-loops with magnetic bases has become an inexpensive and reliable de-facto standard in cryo-crystallography (16), the selection and capture of the crystals from the crystallization drop under a microscope, as well as the micromanipulations during cryo-protection and soaking sweeps are still performed manually. Once a crystal has been flash-cooled to cryogenic temperatures, further manipulation of the cryo-pins and the actual placing of the rather sturdy pins onto the goniostat can be performed quite fast and reliably by robotic arms with grippers.

5.2.2.1. Selection and harvesting of crystals

Crystals are grown using a variety of crystallization techniques, and not all techniques are equally well suited for crystal harvesting. Practically all high throughput crystallization experiments are set up in some SBS (Society of Biological Screening) standard compliant multi-well format with 96 to 1536 wells. Initial crystallization screening against many different conditions is usually performed with the objective of minimizing material usage, and suitable micro-batch screening methods under oil (17) or free interface diffusion experiments in micro-chips (18) are not necessarily designed with ease of harvesting in mind. In optimization experiments, where minor variations around successful crystallization conditions are set up, vapor diffusion sitting drops of 1 µl to 50 nl suitable for harvesting are most commonly used. Nano liter drop sizes not only determine the maximum size of crystals that can be obtained, but also significantly affect nucleation and growth kinetics (19). The use of nano liter drop technology was the subject of an infringement dispute involving the patent holder Syrrx in San Diego and Oculus Pharmaceuticals (U.S. Patent 6,296,673 "Methods and apparatus for performing array microcrystallizations")

For harvesting, crystals are selected under the microscope and lifted from the drop with a small, suitably sized cryo-loop. A number of research groups have developed plate scanning and crystal recognition software, which can reliably detect crystals (at least those of reasonably defined shape) and the methods are expected to improve further (20-22). Crystals are judged by size and appearance (often deceiving), and crystals with isotropic dimensions in the range of 100 µm are considered most desirable. The growing availability of powerful and automated micro-focus beam lines on 3rd generation synchrotron sources allows, in ideal cases, very small crystals approaching 10-20 µm in the smallest dimension to be used, thus also permitting successful data collection on highly anisotropic crystal needles or plates (23). Below µm size, intensity issues and line broadening due to limited periodic sampling, as well as radiation damage in the protein crystal, generally become limiting factors (24).

5.2.2.2. Soaking and derivatisation of crystals

For *de novo* phasing methods based on the determination of a marker atom substructure, some atoms in the protein must act as sources of isomorphous and/or anomalous differences (See table 2, Phasing Strategies). Such marker atoms can be natively present heavy atoms, such as Fe, Zn, or Cu. In sulfur single-wavelength anomalous diffraction phasing (S-SAD), the sulfur atoms of Cys or Met residues act as marker atoms. In those cases of native marker atoms, no derivatization or soaking is necessary. The advantage of no need for soaking also holds for Se-Met labeled proteins, and no marker atoms are needed when Molecular Replacement (MR) phasing with homologous models will be attempted.

Non-native marker ions can either be co-crystallized (i.e., added a priori to the crystallization cocktail), or the native crystals can be soaked for minutes to hrs in mother liquor with a metal ion added in mM concentration, followed by short, optional back-soaking to remove the unbound ions from the crystal's solvent channels (25, 26). Soaking and co-crystallization techniques are also used to incorporate ligands, cofactors, inhibitors or drugs into the crystals.

Heavy metal ions, or halide anions such as bromide and iodide (27, 28) can also be introduced during brief sweeps in combined heavy ion - cryo-buffers. Due to the location of metal or iodine X-ray absorption L-edges (or even uranium M-edges) not too far below the characteristic Cu-K_a wavelength (8keV), SAD/SIRAS phasing should become an increasingly interesting (in-house) alternative to synchrotron based (multi-wavelength) methods (27).

5.2.2.3. Cryo-protection and loop mounting

Rapid cooling of crystals to cryogenic temperatures (quenching or flash-cooling[2]) has become a standard procedure in macromolecular crystallography (29). The foremost reason for cryo-cooling is the drastic reduction of radiation damage (24), eliminating the

[2] The term freezing should be avoided as it encompasses the definition of ice formation, which is detrimental to the crystals.

need for multiple crystals with the associated data merging errors; and secondly increasing the resolution due to reduced thermal vibrations. An additional benefit is that, once the crystals are cryo-protected and safely embedded in the solid amorphous mother liquor of their loops, they can be handled quite easily and reliably by high throughput mounting robots.

Successful cryo-cooling depends on a number of factors, only few of them well established under controlled conditions. If the crystallization cocktail does not *a priori* contain high enough concentrations of reagents like PEG, MPD, or glycerol to prevent freezing (i.e., the formation of ice destroying the crystal), the mounted crystal needs to be swiped through a cryoprotectant before being flash cooled. Cryoprotectants prevent ice formation in the mother liquor, which can be established by quenching an empty loop in liquid nitrogen or the diffractometer nitrogen cold stream and checking the diffraction pattern for the absence of ice rings (a diffraction pattern displaying typical ice rings is shown in (30)). For reasons not entirely clear, even rapid quenching and amorphous state of mother liquor do not guarantee successful cryo-cooling. Excessive increase in mosaicity and loss of resolution are common mishaps. A number of annealing procedures occasionally reducing mosaicity have been reported (31, 32), but these have not yet become established standard procedure in HTPX pipelines. Although a common recipe for cryoprotectants is to spike mother liquor with glycerol, PEG, MPD or other additives, few systematic studies of generally applicable procedures for high throughput efforts appear in the literature. Clearly, these methods are of great relevance to the objectives of high throughput crystallization, but they are time consuming and risky for the crystals, with little systematic or automated procedures developed so far. Problems during soaking and cryo-protection are probably the single most significant source of loss of crystals in both high and low throughput crystallography.

5.2.2.4. Robotic sample mounting

Automated mounting of the cryo-pins on the diffractometer greatly enhances utilization of valuable synchrotron beam time. Practically every major synchrotron facility and larger niotech companies have developed mounting robots for their HTPX beam lines (see for example, (33, 34), Table 1). Commercial systems which are also

suitable for in-house lab sources, are becoming available (Mar Research, Rigaku-MSC, Bruker-Nonius). Under the premise that every crystal deserves screening, fast and reliable storage and mounting procedures are needed to achieve high-throughput data collection. The sample transport and storage system developed at the at the Advanced Light Source (ALS) Macromolecular Crystallization Facility together with the Engineering Division of Lawrence Berkeley National Laboratory (35) may serve as an example for an easy-to-use and quite practical development. The basic handling unit, a cylindrical, puck-shaped cassette containing 16 cryo-pins, also serves as an integral part of a complete, automated cryogenic sample alignment and mounting system, which has been routinely operating since 2002 at the ALS protein crystallography beam line 5.0.3 (Figure 2).

Seven puck cassettes, each holding sixteen Hampton-style, magnetic base cryo-pins, fit into a standard dry shipping dewar. The pucks are loaded at the crystallization lab with the crystals on cryo-pins, and four pucks are transferred into the robot-hutch liquid nitrogen vessel. The mounting robot can randomly access any sample with a cooled, robotic gripper which transfers the sample to the diffractometer within seconds, maintaining the crystal temperature below 110 K. Crystals are centered automatically on the computer controlled goniometer head. In addition to saving valuable beam time, mounting robotics also allow the safe removal and re-storage of a sample, should the initial analysis of diffraction snap-shots cast doubt on the crystal quality. Potentially better crystals can be mounted and examined without risk of loosing the best one found so far.

5.2.3. Data collection

Data collection is in fact the last physical experiment that is conducted in the long process of a structure determination, and it deserves full attention. Considering the constant loss of targets throughout the steps of expression, purification, and crystallization, failure to obtain useful data in the final experiment is most costly. Diffraction data quality largely (and without mercy) determines the quality, and hence usefulness and value, of the final protein structure model.

5.2.3.1. High throughput considerations

In high throughput mode it may not be worthwhile, except in special cases, to collect data sets with resolution worse than 2.3-2.5 Å, but better to pursue additional crystallization optimization or protein engineering. Overall throughput might well be higher when adopting a high resolution strategy, particularly in view of the increased difficulty to accurately build and refine low resolution models, and considering the reduced information content in low resolution models. A decision whether to pursue a low resolution structure will be influenced by whether a structure serves the purpose of fold determination, or must satisfy the more stringent quality criteria for a drug target structure.

5.2.3.2. Data collection as a multi-level decision processes

Once the cryo-pins with the crystals are transferred to the robot-hutch dewar, the remaining steps of the crystal structure determination can principally proceed in fully automated mode. At present however, there are still weaknesses and substantial off-line processing in the data collection stage. At several points, strategic decisions need to be made whether to accept a given level of data quality (and hence, a certain probability of failure in the structure determination), or to proceed to the next crystal. Clearly, reliable robotics provide the advantage of safely un-mounting and storing an acceptable but not optimal crystal for later use, and to proceed to evaluate hopefully better ones.

Any expert system handling the chain of decisions during initial crystal assessment and data collection must be able to evaluate a significant number of parameters at its decision points (36). Problems and irregularities can occur at several stages during data collection, and often show up and become critical only later on. At present, data collection expert systems are not yet developed to completely handle the entire decision process. Such a system requires tight interfacing with the data indexing, integration and reduction programs, and with the beam line hardware and robotics control software (reference (37) exemplifies the substantial complexities involved).

5.2.3.3. Initial assessment of crystal quality and indexing

The most prevalent data collection technique in protein crystallography, practically exclusively used in HTPX efforts, is the rotation method. During each exposure, the crystal is successively rotated by a small increment (usually 0.2 - 1.0 deg) around a single axis, until sufficient coverage of the reciprocal (diffraction) space unit cell is attained.

Once the crystals are centered on the goniostat, a first diffraction pattern is recorded. The crystal is exposed to a collimated, fine beam of X-rays - for a few seconds on a powerful synchrotron, and up to several minutes on weaker or laboratory X-ray sources. The diffracted X-rays are recorded mostly on CCD area detectors (longer read-out times disfavor image plate detectors for high throughput use on synchrotrons), and recorded as an image of diffraction spots, also referred to as a (rotation or oscillation) frame. The first frame immediately shows the extent to with the crystal diffracts. Good diffraction implies single, resolved, and strong spots, extending far out in diffraction angle to high resolution. The relation between diffraction, resolution, and structure quality is shown in figure 3. The first diffraction snapshot also can reveal the presence of ice rings (38). Although icing affects the reflections in proximity to the ice ring, frames with not too excessive ice rings can be processed with little difficulty (29, 39).

Depending on the quality of the data and the indexing algorithm used, it can be possible to index the diffraction pattern based on a single frame or snapshot (39). Indexing means the assignment of a consistent set of three reciprocal basis vectors, which span the reciprocal lattice represented by the diffraction spots. The corresponding direct vectors (a,b,c) and angles between them (a,ß,?) define the crystal unit cell. In practice, more than one frame is used for indexing, for several reasons: Crystals may not diffract isotropically, and snapshots in different orientations assure that anisotropy does not cause unacceptably low resolution in certain directions/orientations of the crystal. A single frame also may not contain enough reflections to allow reliable determination of the internal Laue symmetry of the diffraction pattern, which again determines the possible

Laue group and crystal system of the crystal[3]. Several space groups may be possible under each Laue group, and it is not always possible to unambiguously determine the crystal's space group at this early stage from systematic absences of reflections alone. Proper determination of the Laue group is necessary to develop a strategy to collect a complete set of diffraction data. The data collection strategy also depends on the selected phasing strategy, as discussed below.

Typical difficulties arise during indexing, when it is not possible to find a consistent unit cell for the crystal. Large mosaic spread, large spot size, streaking and overlap, multiple or satellite spots due to macroscopic twinning, and excessive ice rings can cause problems. Spot overlap due to large unit cell dimensions should automatically trigger a re-evaluation at larger detector distances, and a strategy with multiple sweeps at increasing detector offset angles may have to be generated (discussed also under ultra high resolution strategies). After indexing, the choice of Laue symmetry may not be unambiguous, and if in doubt, a lower symmetry must be selected in developing the data collection strategy, depending on the planned phasing technique. Even after successful data reduction and space group determination, the possibility of microscopic twinning, not recognizable from the appearance of the diffraction pattern, exists in certain space groups, and should be automatically evaluated (40).

Increased frame exposure time in pursuit of high resolution tends to lead to low resolution detector pixel saturation. Automatic detection of saturation should routinely trigger collection of a second, faster low resolution data collection sweep. The need for good low resolution data for any phasing method (including MR) has been pointed out repeatedly (41).

5.2.3.4. Data collection strategies for phasing

As indicated in Figure 1, each phasing technique requires a suitable data collection strategy to obtain the necessary coverage of the reciprocal (diffraction) space.

---

[3] It is a common misconception that the crystal system is determined by the cell constants and angles. The internal symmetry overrides the apparent symmetry deduced from the cell constants. It is possible, for example, that an apparently orthorhombic cell ($a \neq b \neq c$, $a=ß=?=90$ deg) is monoclinic, with $ß=90$ deg.

It is seldom a disadvantage to collect as much redundant data as possible, and given the high throughput capabilities of synchrotrons, a few general strategies suffice to cover most standard phasing techniques (42, 43).

The simplest case of data collection is a single wavelength data set without consideration of the anomalous signal. Although anomalous contributions from all atoms in the crystal are present at varying degrees at all wavelengths, they are miniscule for the light elements (H,C,N,O) comprising most of the scattering matter in native proteins, and special techniques described later are employed to utilize the minute anomalous signal of sulfur for phasing. A single wavelength data set covering the reciprocal space unit cell contains the data necessary for structure solution by Molecular Replacement (MR). Except in cases of special orientation of a crystal axis nearly parallel to the rotation axis in uniaxial systems, and at very high resolution, a practically complete set of data can be collected in a single sweep of successive frames (42). The extent of the necessary rotation range is calculated by a strategy generator from the crystal orientation matrix, instrument parameters, and the Laue symmetry. As always, excessive pixel saturation of intense low resolution reflections may require a second sweep with shorter exposure times.

5.2.3.5. Data collection for high resolution structures

In fortunate cases, crystals diffract to very high resolution. The term is loosely used, and shall indicate in our case crystals diffracting better than about 1.5 Å, with exceeding 1.2 Å denoted as the onset of atomic resolution, or 'ultra high' resolution. As illustrated in Figure 3, high resolution permits to discern fine details in the structure, which can be understood as a manifestation of tighter sampling intervals or 'slices' throughout the crystal. As the number of reflection increases with the volume of sampling space, even a numerically less impressive increase in resolution leads to a large increase in recorded data (figure 3), thus drastically improving the accuracy and precision of the subsequent structure refinement. Given the (rare) case of very strong data with resolution of at least 1.2 Å (Sheldrick's Rule, (44, 45)) and small protein size, structures can be determined *ab initio* via Direct Methods (section 5.2.4.2, (46)).

Additional effort is required to collect complete data sets at very high resolution. In certain crystal orientations, it is principally impossible for geometrical reasons to

record all reflections. A part of the diffraction pattern (affectionately called the 'apple core') remains unrecorded. While this range is small (a few %) at 'normal' resolution, it can become large at very high resolution, and suitable hardware that allows movement of the crystal about another axis must be interfaced with the strategy devising program. Despite the large recording area of modern detectors and the short wavelengths used at synchrotrons[4], additional sweeps at larger detector offset angles can become necessary and the data collection strategy quite elaborate, as in the early days of area multi-wire detectors with small solid angle coverage. A finer slicing of the rotation range in frames of about 0.2 deg has certain benefits (47), and an increasing number of data collection programs will probably implement this option in the near future (48).

5.2.3.6. Single anomalous diffraction data and SAD from sulfur

Anomalous data collection requires that in addition to a unique wedge of data covering the reciprocal space asymmetric unit, the Friedel mates (reflections of inverse indices in the centrosymmetrically related part the diffraction pattern) must be recorded. A most useful difference in intensity between Friedel mates results from wavelength dependent, anomalous scattering contributions, and intensity difference data are the basis for location of the anomalously scattering atoms in the phasing stage (see section 5.2.4).

Anomalous data are recorded in smaller blocks (15-30 deg) of data and their inverse segment. Possible radiation damage or beam decay require this precaution. Splitting the data set into smaller blocks that include the corresponding inverse has the additional benefit that, after recording the first block, the significance of the anomalous difference signal can be determined, and the data collection expert system should adjust data collection times accordingly. Alternatively, if no usable anomalous signal can be

[4] According to Bragg's Law, shorter wavelength (higher energy) 'compresses' the diffraction pattern and more data can be recorded within the same solid angle. Large unit cell dimensions, which require larger crystal-detector distance for spot resolution, can partly eliminate this benefit.

expected in reasonable time[5], it may be more efficient to abandon data collection and to proceed to another crystal.

Anomalous data collection for single wavelength experiments does not require measurement of an X-ray absorption spectrum (XAS, Figure 4). The experiment must be conducted at or above the absorption edge of the selected marker element, but the exact determination of the absorption edge spectrum, as is necessary to optimize dispersive ratios in MAD experiments, is not needed. In cases where 'white lines' in the spectrum can be present (elements of the 3rd period and higher), experimentally determining the exact absorption maximum is of advantage for maximizing the anomalous differences.

Special considerations are required when using native sulfur of the Met and Cys residues as anomalous marker. Even the longest practically usable X-ray wavelengths[6] (around 2 Å) are far above the K absorption edge of sulfur, and the anomalous difference signal is often as low as ~ 0.5 % of the total signal (49). However, given sufficiently redundant data collection via integration of multiple sweeps covering the reciprocal space unit many times (720 deg and more of rotation and varying crystal orientation), data with S/N ratio sufficiently high to extract anomalous intensity differences can be collected. In combination with powerful density modification techniques, the SAS method proposed 20 years ago by Wang (50), has recently been shown to be quite successful (49, 51). Given that no special marker atoms need to be introduced into the protein, the method is likely to gain rapid acceptance in high throughout crystallography (see Matthews (52) for a review about SAD data collection and phasing).

---

[5] Note that to obtain twice (n=2) the S/N ratio or signal, four times (n^2) as much data collection time is needed, which over-proportionally reduces throughput and hence, process efficiency.

[6] Geometric and X-ray optical constraints of the tunable beam line components, as well as rapidly increasing absorption at longer wavelengths (lower X-ray energies) set a limit to how long a wavelength can be experimentally used.

5.2.3.7. Multiple anomalous diffraction data

MAD phasing (53) exploits additional redundancy in anomalous signal by not only using anomalous (Friedel- or Bijvoet-) differences *within* each data set, but also dispersive differences *between* data sets recorded at different wavelengths. To optimize these differences, an accurate experimental absorption edge scan for the phasing element needs to be recorded and between two and four MAD wavelengths are selected (Figure 5). The anomalous differences are largest at the absorption edge maximum (max f''), and the dispersive differences are largest between the other data sets and the data set recorded at the inflection point of the edge jump, corresponding to a minimum in f'. The high-remote data are usually recorded several hundred eV above the edge and contain still substantial internal anomalous signal. The least anomalous difference signal, but still dispersive contributions, are expected from the optional low-remote data set, collected several hundred eV below the absorption edge. In view of signal loss through radiation damage, the most common strategy is to obtain peak wavelength data first (still enabling SAD phasing with a reasonable chance), followed by inflection point data (this second data set providing mostly dispersive differences), and the high-remote data set (with redundant internal anomalous differences and large dispersive differences against inflection data). All other basic data collection strategy considerations regarding completeness and S/N ratio discussed in the previous sections apply to MAD data as well. MAD data collection is currently the phasing method of choice in high throughput protein crystallography (12).

5.2.3.8. Raw data warehousing

On top of the data collection and decision making involved to this point, the challenge of raw data warehousing is substantial in HTPX efforts. On high throughput beam lines equipped with 3x3 module CCD detectors, saving a single image (frame) can require over 10 MB of disk space, and data accumulation rates can approach gigabytes per minute. On-line data reduction while rapidly archiving the original image data places substantial demands on the IT infrastructure.

5.2.4. Crystallographic computing

Once the data collection is successfully finished, the remaining steps of the structure solution are carried out *in silico*. Crystallographic computing has made substantial progress, largely due to abundant and cheap high performance computing. It is now (June 2003) possible to solve and analyze complex crystal structures entirely on $2k laptop computers. Consequently, automation has reached a high degree level of sophistication (although many compatibility and integration issues remain). As a result, the actual process of structure solution, although the theoretically most sophisticated part in a structure determination, is commonly not considered a bottleneck in HTPX projects. Given reliable data of decent resolution (~2.5 Å or better) and no overly large or complex molecules, many structures can in fact be solved *de novo* and refined within several hours.

5.2.4.1. Data reduction and scaling

The raw data obtained from the data collection program or expert system need to be further merged, sorted and reduced into a unique set according to the Laue group, and if not already known, the possible space groups need to be determined. If multiple data sets are used for phasing, these data sets must be brought onto a common scale as well. Depending on the amount of integration, this may be handled by the data collection experts system or by sequential programs, and partly by phasing programs. At this stage, the possible number of molecular subunits in the asymmetric unit of the crystal can be estimated, but as in case of the space group, the answer may not be unambiguous and must await the metal substructure solution. An automated HTPX program package or system has to successfully handle the multiple possibilities and provide proper decision branching (36). For isomorphous data sets from different crystals, additional complications resulting from multiple indexing possibilities also need to be expected and accounted for.

5.2.4.2. Phasing - general remarks

The core of the phase problem, which makes protein crystallography non-trivial, is the absence of direct phase information in diffraction data. Two quantities per reflection need to be known in order to reconstruct the electron density (Figure 6): the magnitude of the scattering vectors (or structure factor), which is proportional to the square root the measured reflection intensity, and the relative phase angle of each scattering vector, which cannot be directly measured. Unfortunately, the phases dominate the electron density reconstruction, a fact giving rise to the phenomenon of phase- or model-bias. Incorrect phases from a model tend to reproduce incorrect model features despite experimental data from the true structure. Bias minimization will be discussed in the section about molecular replacement.

Practically all macromolecular phasing techniques used in HTPX depend on the presence and solution of a marker atom substructure (Table 2). By creating difference intensities between data sets with and without the contributions from the marker atoms, the initial problem is reduced to solving a substructure of a few to a few hundred, versus many thousands to ten thousands of atoms (Figure 7). The intensity differences can arise between absence (native) and presence (derivative) of heavy atoms, which forms the basis of isomorphous replacement techniques. Differences can also arise from different anomalous intensities at a single wavelength originating from native (S, Fe, Cu), engineered (Se), or derivative anomalous scatterers (SAD); or from additional dispersive differences between data sets recorded at different wavelengths (MAD)[7]. In anomalous methods, all data are preferably collected from one single crystal and are thus perfectly isomorphous. Combined with highly redundant data collection, excellent experimental phases can be obtained even for weaker high resolution reflections *via* anomalous phasing techniques. Consequently, anomalous methods are the workhorse of high

[7] The pseudo-SIRAS-like treatment of MAD data as presented here and used for example in SOLVE (Terwilliger et al., 2001), is different in details from the explicit solution of the MAD equations originally used by Hendrickson (Hendrickson, 1991).

throughput phasing. Combinations of various phasing techniques are also possible, providing higher redundancy in the phase angle determination.

5.2.4.3. Substructure solution

The solution of the marker substructure is the first step in determination of the phases. Three-dimensional maps containing peaks at the position of interatomic distance vectors between the marker atoms can be created from difference intensity data without the use of phases. Such difference Patterson maps (Figure 8) contain strong peaks in certain sections and along certain directions, and the correct marker atom positions giving rise to a consistent peak pattern can be determined. A number of software packages used in HTPX use Patterson techniques in varying flavors to find consistent solutions in one or more difference maps (54). Due to the centrosymmetry of the Patterson space, the handedness of the metal substructure cannot be determined by Patterson methods alone. Anomalous contributions, direct methods (discussed below), or map interpretability yield additional information to break the inherent substructure enatiomer ambiguity.

Direct Methods provide an alternative avenue of solving the metal substructure *ab initio* from intensities reduced to normalized structure factor amplitudes. Statistical inferences about phase relations of improved starting atom sets from Patterson superposition techniques cycled with real space phase expansion (46, 55) and dual space methods (SnB, (56)) are particularly successful. Substructure solution is highly automated, and Se substructures containing up to 160 atoms have been successfully solved with both SHELXD (57) and SnB (58). In rare cases of strong data, atomic resolution (better than 1.2 Å) and modest size (from few hundred to currently over one thousand non-hydrogen atoms) complete protein structures can be solved by direct methods (46). Direct methods can also extract additional information from intensities determining the absolute handedness of the metal substructure (59). The heavy atom positions are also used to determine the NCS operator needed for subsequent map averaging and density modification (discussed below).

5.2.4.4. Initial phase calculation

The phase angles of the reflections can be determined once positions of the marker atoms are refined, and hence, the magnitude and the phase angle of the marker contribution to the total diffraction intensity are known. Figure 8B shows a graphic representation (Harker diagram) visualizing the solution of the phasing equations. Two limitations become immediately clear. First, the solution is not unique if only one difference data set is available, leading to the need to break a second phase ambiguity in addition to the metal substructure handedness discussed above. Second, based on errors in both the measured intensities and the marker atom positions, each circle intersection defining a phase angle will contain a certain error. The inherent phase ambiguity can be removed by use of multiple derivatives, adding anomalous signal in single derivative cases, or via multiple wavelength methods. Multiple determinations of each phase angle also increase the probability for the phase angle to be correct, and thus increase the figure of merit for the best phase estimate.

Once a phase angle for each reflection is obtained, a first electron density map can be computed. If the handedness of the metal substructure has not yet been determined (except in pure MIR cases), a map containing anomalous contributions will be not or much less interpretable (lower figure of merit) if the wrong enantiomer of the substructure was used in the phasing calculations. In the case of single anomalous phasing data (SAD), even a proper map will contain density of the molecule, superimposed with noise features. To improve the interpretability of all maps and to compensate for the lack of a unique sign of the phases in the SAD case, iterative density modification and filtering techniques (50) are applied in the next step.

5.2.4.5. Density modification techniques

One of the most powerful tools at hand to obtain readily interpretable maps, which is particularly important for automated model building, are (direct space) density modification techniques. Implementations of solvent flattening (60), solvent flipping (61), histogram matching (62) and reciprocal space maximum likelihood methods (63) exploit the fact that protein molecules pack loosely in the lattice. Substantial solvent

channels (about 30 to 70 % of the crystal volume) are filled with non ordered solvent, giving rise to a uniform density distribution in the solvent region. Setting the solvent electron density to a constant value ('flattening'), in repeated cycles with adjustments in the solvent mask under extension of resolution (64), leads to drastically reduced phase angle errors, and hence, clearer and better interpretable electron density maps (Figure 9). Density modification is also important in permitting phase extension to higher resolution in the frequent case where the derivative or anomalous data used for substructure solution and phasing do not extend to the same high resolution as the native data (65).

Another powerful variation of density modification is map averaging, which is applicable both in presence of non crystallographic symmetry (NCS) and in model bias removal techniques based on multiple perturbed models (discussed later). The principle of NCS averaging is that if more than one molecule is present in the asymmetric unit of the crystal due to additional non-crystallographic symmetry, the diffraction pattern and hence the back-transformed map, will contain redundant information. The electron density of the different copies of the molecule can be averaged (consistent features will amplify whereas noise and ambiguous density will be suppressed), and again, a greatly improved electron density map of the molecule results[8]. Map averaging is also possible between different crystal forms of the same protein, and in view of the increasing number of different crystal forms obtained via high throughput crystallization efforts, may be attractive to routinely implement.

Full automation of NCS averaging is not trivial. One subtlety is that the atoms of metal substructure often are found in different asymmetric units and/or in adjacent unit cells. Early determination and refinement of the NCS operators is necessary for subsequent map averaging, as is determination of the proper molecular envelope (or mask) to submit only the map of one complete molecular subunit for initial automated model building. Increased attention towards automated utilization of NCS (66, 67) will eventually lead to stable integrated expert systems handling automated map averaging.

[8] Density modification by map averaging is in fact so powerful that virus capsid structures, which are highly symmetric and contain up to 60 copies per molecule, can be phased without marker techniques.

5.2.4.6. Map interpretation and model building

Once an electron density map of best possible quality is obtained from improved experimental phases, a model of the protein structure must be built into the electron density. The process is generally more successful with clean maps and at higher resolution. Traditionally, model building was carried out by hand using programs that graphically display the electron density and allow placement and manipulation of protein backbone markers and residues, combined with varying real space and geometry refinement tools (Table 3). High throughput requires that the interpretation of the map, building, and if possible, refinement and correction steps are carried out automatically by specialized programs without the need for graphical user interaction. The programs generally follow a procedure similar to what is used in manual model building, with the benefit of fast and automated library searches. In the first step, Ca backbone atoms are placed at recognizable branching points of main and side chain density, and the longest contiguous chain is sought (68, 69). Search of fragment libraries (60) and use of preferred rotamers (70) can improve the model quality already in the first building cycles. At higher resolution, the electron densities of the residues are also less degenerate than at low resolution, aiding proper feature recognition and faster sequence alignment. Sequence anchoring either on stretches of distinct residues or at marker atom sites facilitates tracing of the chains properly as well as to recognition and correction of branching errors. Interestingly enough, no automated model building program currently appears to take advantage of the simultaneous positional *and* sequential marking by Se atoms. One automated program, RESOLVE (71), uses iterative model building in combination with maximum likelihood density modification, and ARP/wARP is based on cycled dummy atom and fragment building and refinement (72). Table 3 contains a summary of additional public and commercial programs. A brief direct comparison of RESOLVE, MAID (73) and ARP/wARP has been compiled recently (74).

5.2.4.7. Refinement

The initial raw model built into the electron density must be refined, and by phase combination of experimental and model phases, an improved map is obtained. The model

building step is repeated, additional model parts are built into the improved density and necessary corrections are made to the model. The refinement itself consists of adjusting some general scaling parameters to match observed and calculated data overall, and the atomic coordinates of the initial model are refined so that the differences between observed and calculated data are minimized (75). The common global measure of this agreement is the R-value, often expressed in %.

$$R = \frac{\sum_{hkl} \left| |F_{obs}| - k|F_{cal}| \right|}{\sum_{hkl} |F_{obs}|}$$

With improving model quality, individual atomic temperature factors (a measure for positional displacement either via thermal motion or through disorder) are also refined, and in cases of atomic resolution, it may be possible to refine ansiotropic temperature factors. Even at modest resolution, however, grouping and refining parts of the molecule through TLS (torsion, libration, screw) motion modes improves refinement and model quality (76). Bulk solvent model corrections are also applied during the refinement (77).

A general problem in protein structure refinement is the low data/parameter ratio. Such refinements are not stable against experimental data alone, and additional restraint terms creating penalties for deviations from geometry target values are used (78). The poor data/parameter ratio not only requires special refinement techniques, but also implementation of safeguards against overfitting and introduction of model bias. Strict crossvalidation against a small subset of unused (free) data, monitored by the R value for the free data set (79) is standard practice. Properly used, this free R prevents overfitting, i.e., introduction or variation of parameters which do not contribute to model improvement but only reduce the fitting residual (an example would be excessive solvent building). Maximum Likelihood (ML) refinement target functions (80) allowing for error in the model based on the sigmaA map coefficients (81) are now universally implemented and reduce, but not eliminate, susceptibility to model bias. Refinements using ML targets and torsion angle refinement combined with simulated annealing techniques (82, 83) have a large convergence radius and are especially useful for refinement of initial models far from the correct one, as is often the case in Molecular Replacement (discussed below). Despite substantial progress in automated model

building, a surprisingly large number of final adjustments and repair of house-keeping errors still need to be made manually to fine-tune the structural model. Integration and cycling of refinement and building with real time validation, including local validation via real space fit correlation and chemical plausibility, will further improve the quality of automatically built and refined models.

5.2.4.8. Automated molecular replacement

The principle of MR is to use a homologous search model, sufficiently close to the unknown structure, and to properly 'replace' - in the sense of 'repositioning' - the molecule(s) in the unit cell until good correlation between the observed diffraction data and the data calculated using the replaced search model indicates a solution. Once the correct position of the search model is determined, initial - but highly biased - phases for the unknown structure can be obtained. About 75% of the structures in high throughput efforts are currently solved using MR, particularly in repeated structural screening of the same protein co-crystallized with different drugs. Rapid data collection and automated molecular replacement routines are the backbone of high throughput structure based drug screening (1).

A single native data set only is necessary for MR phasing, provided a homology model within no more than a few Å backbone coordinate r.m.s.d. is found in the Protein Data Bank (PDB, (84)), or built by computer modeling (85). Generally, accuracy of the model appears to be more important than completeness for convergence of the MR search (86). Conventional search programs perform separate rotational and translational searches to find the proper position. Innovations in the method such as full 6-dimensional, fast evolutionary searches (87) or combinations with new maximum likelihood based approaches increasing the radius of convergence of the searches, are becoming established (83, 88). The process is easily automated, and given the anticipated rise in coverage of structural folds available in the public databases due to the structural genomics efforts, MR will very likely see constantly increasing use.

A general strategy for automated search model building is to identify a set of possible template structures with sequence alignment tools and to retrieve them automatically from the protein structure database. Search models are built from each

template either directly or obtained via homology modeling, and target side chains can be automatically built. Parallel molecular replacement searches for each of the highest scoring models are branched to a computer cluster and the models are evaluated according to their correlation coefficient to observed data. Fold recognition models, although steadily increasing in quality (85), still may not produce successful MR probes. The immediate feedback possible through evaluation of the model against experimental data, however, should allow for adaptive correction of the model building algorithms in response to MR scoring. Model completion techniques such as loop building and gap filling are likely to benefit from such experimental restraints.

A drawback of the MR method is its high susceptibility to model (phase) bias recreating the model's features in the electron density, and the bias minimization techniques described in the refinement section must be rigorously applied. Effects of model bias can be insidious (89) and are not easily recognized by commonly used global structure quality descriptors such as R and free R (90). A fully automated model bias removal protocol (Shake&wARP, (91)), based on a modified dummy atom placement and refinement and protocol (92) uses a combination of model perturbation, omit techniques, and maximum likelihood refinement together with multiple map averaging techniques to effectively minimize phase bias.

5.2.4.9. Initial model validation

The refined final model is subjected to an array of validation techniques, ranging from a ranking *via* global validators against other structures to detailed chemical and folding plausibility checks based on prior knowledge (Chapter 10). Global indicators such as R and even the cross-validated free R value cannot be specific as far as the local quality of a structure model is concerned. Similar considerations hold for average deviations from geometry targets, which largely reflect the weight of the restraints chosen in refinement. While limited local errors may not be a great concern for a structure solved to populate fold space[9], a drug target structure needs to be of high quality around

[9] Note however, that each and every atom's position in the structure contributes to

the specific drug or ligand binding site. Local methods of assessment evaluate the correlation of the model against a bias minimized electron density map (93), or check local geometry descriptors for outliers on a per residue basis (PROCHECK (94), WHAT-CHECK (95), see also Chapter 10). Extended stretches of consistently high deviations in either case are indicative of serious 'problem zones' within the model. An adaptive response by the model building program to such analysis would be desirable in fully automated structure determination packages. Presently, the validation programs are largely self-contained and stand-alone programs, and model corrections are still made off-line and *a posteri.* Correction of nuisance or housekeeping errors such as stray atoms, nomenclature violations, sequence inconsistencies, etc. could be easily automated. Currently, most of these nuisance errors are corrected only at the off-line, PDB deposition level (Autodep, ADIT, Chapter 10). It should be expected that such corrections will be automated to a much higher level, up to fully automated data harvesting and deposition/annotation procedures.

In the author's experience, automatically solved and built structures do not seem to be any less reliable (or more 'wrong') than conventionally determined structures. The automated phasing programs fail to a similar degree as those scripted by a human operator, but automation provides the benefit that a much more rigorous and consistent pursuit of multiple options can be implemented. Model building programs manage fairly well at the "beginner's" level with good maps. They tend to fail in borderline cases when a skilled crystallographer might still be able, with serious effort, to successfully bootstrap a build. On the other hand, building programs are also devoid of any desire to salvage an abysmal project, which avoids a number of bias issues *a priori* (see (89) or (96) for illustrative examples). The author feels that the danger of flooding the structural data bases with low quality models from automated HTPX structure determination projects is overstated.

5.3. Summary of progress and challenges

Development of high throughput techniques for crystallographic structure determination has been fairly rapid, with a much clearer picture of the more challenging areas now emerging. While the technical advances of public efforts are relatively easy to track, proprietary developments are much less accessible, and throughput figures tend towards the optimistic side (97).

5.3.1. Synchrotron x-ray sources

The intensity and brilliance of 3rd generation synchrotron light sources has reached dose limit levels where further increase in brute intensity will not substantially increase throughput (ESRF ID13, APS 19ID). Significant radiation damage to the crystals (98), and recognizable modifications of the protein molecules by high radiation (99) already require already attenuation of the most powerful X-ray beams. Upgrades of older synchrotron sources like SSRL's SPEAR to 3rd generation output levels, however, will add significantly to the available synchrotron capacity. Tunable micro-focus sources based on Compton scattering have potential (100), and future developments may lead to viable instruments and new techniques such as broad bandwidth SAD phasing. The much-discussed free electron lasers will probably not readily contribute to future structural genomics efforts, even if the considerable technical problems can be overcome (101).

5.3.2. Robotic crystal harvesting and mounting

Increasing the capacity of existing beam lines by efficient use of beam time is a key benefit of rapid automated mounting methods. A few $100k invested in robotics can substantially increase the throughput capacity of $10M beam lines. In contrast, crystal harvesting, cryo-protection, soaking and loop-mounting are still largely off-line procedures. The (mis)handling of crystals during these steps is probably the single most significant point of failure in the structure determination process. Whether the manual mounting steps will turn into bona fide throughput-limiting bottlenecks will depend on what level of sustained throughput capacity protein production and crystallization can achieve.

5.3.3. Fully automated data collection systems

Data collection, through its intimate connection of direct experiment control and computation with decisions relatively late in the process, are a challenge for full automation. Increasing sophistication and automation of the upcoming versions of data collection suites (36, 47, 102) and their seamless interfacing into the following phasing programs will likely reduce the number of abandoned data collections and data sets, and hence increase throughput.

5.3.4. Automated phasing, model building, and refinement, and deposition

Cooperation between publicly sponsored program developer teams such as CCP4, SOLVE or PHENIX, as well as improvement and interfacing of independently developed suites such as SHARP or MAID (Table 3) have already greatly increased the ease of the computational parts of structure solution. The next logical step would be to interface tightly with the data collection suites acting as a front end for the structure solution programs, and to incorporate feedback from independent validation programs into model building, refinement and deposition (103) of validated models. Although in each of these areas excellent programs exist, seamless integration is still missing.

5.3.5. Interfacing with Structural Bioinformatics

As already discussed in section 5.1.3., structures are determined in HTPX efforts with mainly two objectives in mind: Use of the structure to determine a potentially new fold of unknown function, or as a lead structure for drug design. Although the former is probably not the real high throughput driver, direct interfacing with the fold comparison programs such as DALI (104) and further fold analysis including active site searches (see chapter 19 of (105)) would be desirable. Data harvesting, automated annotation (106), and interfacing to laboratory management systems are just some aspects of integration within the realm of structural bioinformatics (105).

As structure guided drug design is perhaps the major driver for HTPX (1), automated ligand building and docking have been the focus of development in commercial software (for example Accelrys.com, Astex.com), and these approaches are currently being implemented in at least one publicly available model building program (warpNtrace, Victor Lamzin, EMBL Hamburg, personal communication). Interfacing to Virtual Ligand Screening, automated lead optimization, and in-silico ADMET property prediction programs would be the next step towards fully automated drug HTPX guided target structure analysis.

5.3.6. Throughput estimates

A quick analysis of the PDB deposition data reveals that the common notation of an exponential increase in PDB structure depositions per year cannot be maintained. After a brief, nearly overexponential surge in the early 1990 (perhaps largely to the credit of Hendrickson's MAD technique) the number of depositions per year has increased less rapidly, and the curve has flattened considerably since the late 1990s. If the deposition rate indeed reflects the impact of new technology, then one would expect a similar deposition surge in the near future, similar to that occurring in the early 1990s, as a result of the HTPX efforts. The question arises to what degree the synchrotron sources (and improved anomalous in-house phasing techniques) then can satisfy the need for more and more beam time. Assuming a conservative 8 hrs data collection time for complete 4-wavelengths MAD data, and 150 full operating days a year, about 8 such beam lines could produce the data for all structures deposited in one year. As there are approximately 10-15 times as many PX beam lines available throughout the world, a shortage of beam time cannot be easily explained by a lack of hardware. Even accounting for industrial efforts and additional data sets, long waiting times for beam line appear surprising. Suboptimal use of beam time, dropped crystals, aborted scans, unprocessible datasets and unsolved phasing probably account for most of the discrepancy. In nearly all of these instances of failure, intelligent automation (which is not necessarily an oxymoron) will greatly enhance the success rates and efficiency, and allow for a further,

manifold increase in structure determinations and depositions. Nonetheless, a 'killer application' in protein production and crystallization could prove this extrapolation dreadfully wrong.

Acknowledgements:

The following individuals have provided comments, insight, information or updates helpful during the preparation of this chapter: Gerry McDermott, Thomas Earnest, Harry Powell, Gerard Bricogne, Victor Lamzin, Clemens Vonrhein, Dirk Kostrewa, Tim Harris, Aled Edwards, Andrzej Joachimiak, Ehmke Pohl, Martin Walsh, Sean McSweeney, and Katherine Kantardjieff. The author thanks James C. Sacchettini, Texas A&M University, for support of his sabbatical leave at Texas A&M University. LLNL is operated by University of California for the US DOE under contract W-7405-ENG-48. This work was funded by NIH P50 GM62410 (TB Structural Genomics) center grant and the Robert A. Welch Foundation.

5.4. References

5.4.1. General References, Crystallographic Techniques

G Rhodes. Crystallography Made Crystal Clear. 2nd ed. London, UK: Academic Press, 2000. Easy introductory reading for the user of protein structure models

J Drenth. Principles of Protein X-ray Crystallography. 2nd ed. Springer Advanced Texts in Chemistry. New York, NY: Springer, 1999. More detailed derivations requiring some mathematical background.

C Jones, B Mulloy, M Sanderson, eds. Crystallographic Methods and Protocols.Totowa, NJ: Humana Press, 1996. Selected chapters for a quick overview on specific topics

C Giacovazzo, H Monaco, D Viterbo, F Scordari, G Gilli, G Zanotti, M Catti. Fundamentals of Crystallography. 2nd ed. IUCr Texts on Crystallography, Vol. 7. Oxford, UK: Oxford Science Publications, 2002. General crystallographic treatise with emphasis on computational crystallography.

M Rossmann, E Arnold. Crystallography of Biological Macromolecules. International Tables for Crystallography, Vol. F. Dordrecht, NL: Kluwer Academic Publishing, 2001. Concise advanced reading from protein expression to model analysis. Several chapters from this volume are explicitly cited in the special references.

C Carter, R Sweet, eds. Macromolecular Crystallography. Methods in Enzymology, Vol. 276, 277. London, UK: Academic Press, 1997. Collection of in-depth chapters detailing key topics

H Wyckoff, C Hirs, S Timasheff, eds. Diffraction Methods for Biological Macromolecules. Methods in Enzymology, Vol. 114, 115. London, UK: Academic Press, 1985. Overview of state of the art 20 years ago, but still valuable reading for selected chapters.

5.4.2. Complete Journal Special Issues

Nature Structural Biology 5, August 1998, Synchrotron Supplement.

Nature Structural Biology 7, November 2000, Structural Genomics Supplement.

Accounts of Chemical Research 36, March 2003, Special Issue on Structural Genomics.

Journal of Structural Biology 142, April 2003, Macromolecular Crystallization in the Structural Genomics Era.

Acta Crystallographica D58(11), Proceedings of CCP4 Study Weekend on High Throughput Structure Determination.

Acta Crystallographica D57(10), Proceedings of CCP4 Study Weekend on Molecular replacement and its relatives.

Acta Crystallographica D55(10), Proceedings of CCP4 Study Weekend on Data Collection and Processing.

5.4.3. Specific references

1. TL Blundell, H Jhoti, C Abell. High-Throughput Crystallography for Lead Discovery in Drug Design. Nature Reviews Drug Discovery 1:45-54, 2001.
2. N Ban, P Nissen, J Hansen, PB Moore, TA Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science 289:905-20, 2000.
3. I Xenarios, D Eisenberg. Protein Interaction Databases. Curr Opinon Struct Biol 12:334-339, 2001.
4. M Etter. NMR and X-Ray Crystallography: Interfaces and Challenges. ACA Transactions 24, 1988.
5. K Moffat. Ultrafast time resolved crystallography. Nature Struct Biol Suppl 5:641-642, 1998.

6. J Knowles, G Gromo. Target Selection in Drug Discovery. Nature Reviews Drug Targets 2:63-69, 2003.

7. JC Norvell, A Zapp-Machalek. Structural genomics programs at the US National Institute of General Medical Sciences. Nature Struct Biol Suppl 7:931, 2000.

8. B Rupp. High throughput crystallography at an affordable cost: The TB Structural Genomics Consortium crystallization facility. Acc Chem Res 36:173-181, 2003.

9. S Dry, S McCarthy, T Harris. Structural Genomics in the Biotechnology Sector. Nature Struct Biol Suppl 7:946-949, 2000.

10. KE Goodwill, MG Tennant, RC Stevens. High-throughput x-ray crystallography for structure based drug design. Drug Discovery Today 6(15):S113-S118, 2001.

11. U Heinemann, G Illing, H Oschkinat. High-throughput three-dimensional protein structure determination. Curr Opin Biotechnol 12(4):348-354, 2001.

12. W Hendrickson. Synchrotron Crystallography. Trends Biochem Sci 25:637-643, 2000.

13. CM Ogata. MAD phasing grows up. Nature Struct Biol Suppl 5:638-640, 1998.

14. T Harris. The commercial use of structural genomics. Drug Discovery Today 6(22):1148, 2001.

15. GE Moore. Cramming more components onto integrated circuits. Electronics 8(15):2-5, 1965.

16. DW Rodgers. Cryocrystallography techniques and devices. International Tables for Crystallography F:202-208, 2001.

17. A D'Arcy, A MacSweeney, M Stihle, A Haber. The advantages of using a modified microbatch method for rapid screening of protein crystallization conditions. Acta Crystallogr D59:396-399, 2003.

18. C Hansen, E Skordalakes, J Berger, S Quake. A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion. Proc Natl Acad Sci USA 99:16531-16536, 2002.

19. E Bodenstaff, F Hoedemaker, E Kuil, H deVrind, J Abrahams. The prospects of nanocrystallography. Acta Crystallogr D59:1901-1906, 2002.

20. G Spraggon, SA Lesley, A Kreusch, JP Priestle. Computational analysis of crystallization trials. Acta Crystallogr D58(11):1915-1923, 2002.

21. JR Luft, RJ Collins, NA Fehrman, AM Lauricella, CK Veatch, GT DeTitta. A deliberate approach to screening for initial crystallization conditions of biological macromolecules. J Struct Biol 142(1):170-179, 2003.

22. J Wilson. Towards the automated evaluation of crystallization trials. Acta Crystallogr D58(11):1907-1914, 2002.

23. S Cusack, H Belrhali, A Bram, M Burghammer, A Perrakis, C Riekel. Small is beautiful: protein micro-crystallography. Nat Struct Biol Suppl 5:634-447, 1998.

24. E Garman, C Nave. Radiation damage to crystalline biological molecules: current view. J Synchrotron Rad 9:327-328, 2002.

25. SA Islam, D Carvin, MJE Sternberg, TL Blundell. HAD, a Data Bank of Heavy-Atom Binding Sites in Protein Crystals: a Resource for Use in Multiple Isomorphous Replacement and Anomalous Scattering. Acta Crystallogr D54:1199-1206, 1998.

26. TJ Boggon, L Shapiro. Screening for phasing atoms in protein crystallography. Structure 7(8):R143-R149, 2000.

27. G Evans, G Bricogne. Triiodide derivatization and combinatorial counter-ion replacement: two methods for enhancing phasing signal using laboratory Cu Kα X-ray equipment. Acta Crystallogr D58:976-991, 2002.

28. RA Nagem, Z Dauter, I Polikarpov. Protein crystal structure solution by fast incorporation of negatively and positively charged anomalous scatterers. Acta Crystallogr D57:996-1002, 2001.

29. E Garman. Cool data: quantity AND quality. Acta Crystallogr D55:1641-1653, 1999.

30. R Sweet. The technology that enables synchrotron structural biology. Nature Struct. Biol. Suppl. 5:654-656, 2000.

31. S Kriminski, CL Caylor, MC Nonato, KD Finkelstein, RE Thorne. Flash-cooling and annealing of protein crystals. Acta Crystallogr D58:459-471, 2002.

32. LB Hanson, CA Schall, GJ Bunick. New techniques in macromolecular cryocrystallography: macromolecular crystal annealing and cryogenic helium. J Struct Biol 142(1):77-87, 2003.

33. WI Karain, GP Bourenkov, H Blume, HD Bartunik. Automated mounting, centering and screening of crystals for high-throughput protein crystallography. Acta Crystallogr D58:1519-22, 2002.

34. SW Muchmore, J Olson, R Jones, J Pan, M Blum, J Greer, SM Merrick, P Magdalinos, VL Nienaber. Automated crystal mounting and data collection for protein crystallography. Structure 8(12):243-246, 2000.

35. G Snell, G Meigs, C Cork, R Nordmeyer, E Cornell, D Yegian, J Jaklevic, J Jin, TE Earnest. Automatic sample mounting and alignment system for macromolecular crystallography at the Advanced Light Source. J Synchrotron Rad:in press, 2002.

36. A Leslie, H Powell, G Winter, O Svensson, D Spruce, S McSweeney, D Love, S Kinder, E Duke, C Nave. Automation of the collection and processing of X-ray diffraction data -- a generic approach. Acta Crystallogr D58:1924-1928, 2002.

37. TM McPhillips, SE McPhillips, H-J Chiu, AE Cohen, AM Deacon, PJ Ellis, E Garman, A Gonzalez, NK Sauter, RP Phizackerley, SM Soltis, P Kuhn. Blu-Ice and the Distributed Control System: software for data acquisition and instrument control at macromolecular crystallography beamlines. J Synchrotron Rad 9:401-406, 2002.

38. RM Sweet. The technology that enables synchrotron structural biology. Nature Struct Biol Suppl 5:654-656, 1998.

39. HR Powell. The Rossmann Fourier autoindexing algorithm in MOSFLM. Acta Crystallogr D55:10690-1695, 1999.

40. TO Yeates. Detecting and overcoming crystal twinning. Meth Enzymol 276:344-58, 1997.

41. Z Dauter, KS Wilson. Principles of monochromatic data collection. International Tables For Crystallography Volume F:177-195, 2001.

42. Z Dauter. Data-collection strategies. Acta Crystallogr D55:1703-1717, 1999.

43. W Minor, D Tomchick, Z Otwinowski. Strategies for macromolecular synchrotron crystallography. Structure Fold Des 8(5):105-110, 2000.

44. GM Sheldrick. Phase annealing in SHELX-90: direct methods for larger structures. Acta Crystallogr A46:467-473, 1990.

45. RJ Morris, G Bricogne. Sheldrick's 1.2Å rule and beyond. Acta Crystallogr D59:615-617, 2003.

46. G Sheldrick, H Hauptman, C Weeks, R Miller, I Uson. Ab initio phasing. International Tables for Crystallography F:333-354, 2001.

47. W Pflugrath. The finer things in X-ray diffraction data collection. Acta Crystallogr D55:1718-1725, 1999.

48. AGW Leslie. Integration of macromolecular diffraction data. Acta Crystallogr D55:1969-1702, 1999.

49. UA Ramagopal, M Dauter, Z Dauter. Phasing on anomalous signal of sulfurs: what is the limit? Acta Crystallogr D59:1020-1027, 2003.

50. BC Wang. Resolution of Phase Ambiguity in Macromolecular Crystallography. Meth Enzymol 115:90-112, 1985.

51. Z Dauter, M Dauter, ED Dodson. Jolly SAD. Acta Crystallogr D58:496-508, 2002.

52. B Matthews. Heavy atom location and phase determination with single wavelength diffraction data. International Tables for Crystallography F:293-298, 2001.

53. W Hendrickson. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science 254:51-58, 1991.

54. TC Terwilliger, J Berendsen. Automated MAD and MIR structure solution. International Tables for Crystallography F:303-309, 2001.

55. I Uson, GM Sheldrick. Advances in direct methods for protein crystallography. Curr Opinion Struct Biol 9:642-648, 1999.

56. CM Weeks, R Miller. The design and implementation of SnB v2-0. J Appl Cryst 32:120-124, 1999.

57. TR Schneider, GM Sheldrick. Substructure solution with SHELXD. Acta Crystallogr D58:1772-1779, 2002.

58. F vonDelft, TL Blundell. The 160 selenium atom substructure of KMPHT. Acta Crystallogr A58:C239, 2002.

59. Q Hao, YX Gu, CD Zheng, HF Fan. OASIS: a computer program for breaking phase ambiguity in one-wavelength anomalous scattering or single isomorphous substitution (replacement) data. J Appl Cryst 33:980-981, 2000.

60. KD Cowtan. Modified phased translation functions and their application to molecular-fragment location. Acta Crystallogr D54:750-756, 1998.

61. JL Abrahams, AGW Leslie. Methods used in the structure determination of the bovine mitochondiral F1 ATPase. Acta Crystallogr D52:30-42, 1996.

62. KYJ Zhang. SQUASH - combining constraints for macromolecular phase refinement and extension. Acta Crystallogr D49:213-222, 1993.

63. TC Terwilliger. Reciprocal space solvent flattening. Acta Crystallogr D55:1863-71, 1999.

64. KD Cowtan, P Main. Phase combination and cross validation in iterated density-modification calculations. Acta Crystallogr D52:43-48, 1996.

65. KYJ Zhang, KD Cowtan, P Main. Phase improvement by iterative density modification. International Tables for Crystallography F:311-324, 2001.

66. TC Terwillger. Statistical density modification with non-crystallographic symmetry. Acta Crystallogr D58:2082-2086, 2002.

67. C Vonrhein, GE Schultz. Locating proper non-crystallographic symmetry in low-resolution electron-density maps with the program GETAX. Acta Crystallorg D55:225-229, 1998.

68. TR Ioerger, JC Sacchettini. Automatic modeling of protein backbones in electron density maps via prediction of C? coordinates. Acta Crystallogr D58:2043-2054, 2002.

69. T Oldfield. Automated tracing of electron density maps of proteins. Acta Crystallogr D59:483-491, 2003.

70. RL Dunbrack, Jr. Rotamer libraries in the 21st century. Curr Opinion Struct Biol 12:431-440, 2002.

71. TC Terwilliger. Maximum likelihood density modification. Acta Crystallogr D56:965-972, 2000.

72. R Morris, A Perrakis, V Lamzin. ARP/wARP's model-building algorithms. I. The main chain. Acta Crystallogr D58:968-75, 2002.

73. DG Levitt. A new software routine that automates the fitting of protein X-ray crystallographic electron-density maps. Acta Crystallogr D57:1013-1019, 2001.

74. J Badger. An evaluation of automated model-building procedures for protein crystallography. Acta Crystallogr D59:823-827, 2003.

75. LF TenEyck, K Watenpaugh. Introduction to refinement. International Tables for Crystallography F:369-374, 2001.

76. MD Winn, MN Isupov, GN Murshudov. Use of TLS parameters to model anisotropic displacements in macromolecular refinement. Acta Crystallogr D57:122-223, 2001.

77. D Kostrewa. Bulk Solvent Correction: Practical Application and Effects in Reciprocal and Real Space. CCP4 Newsletter Protein Crystallogr 34:9-22, 1997.

78. JH Konnert, WA Hendrickson. A restrained-parameter thermal-factor refinement procedure. Acta Crystallogr A36:110-119, 1980.

79. AT Brünger. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. Nature 355:472–475, 1992.

80. GN Murshudov, AA Vagin, ED Dodson. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. Acta Crystallogr D53:240-255, 1997.

81. R Read. Model Phases: Probabilities, bias and maps. International Tables for Crystallography F:325-331, 2002.

82. AT Brünger, J Kuryan, M Karplus. Crystallographic R Factor refinement by molecular dynamics. Science 235:458-460, 1987.

83. P Adams, N Panu, R Read, A Brünger. Extending the limits of molecular replacement through combined simulated annealing and maximum-likelihood refinement. Acta Crystallogr D55:181-190, 1999.

84. HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, PE Bourne. The Protein Data Bank. Nucleic Acids Res 28:235-242, 2000.

85. DT Jones. Evaluating the potential of using fold-recognition models for molecular replacement. Acta Crystallogr D57:1428-1434, 2001.

86. CR Kissinger, DK Gehlhaar, BA Smith, D Bouzida. Molecular replacement by evolutionary search. Acta Crystallogr D57(10):1474-1479, 2001.

87. CR Kissinger, DK Gelhaar, DB Fogel. Rapid automated molecular replacement by evolutionary search. Acta Crystallogr D55:484-491, 1999.

88. R Read. Pushing the boundaries of molecular replacement with maximum likelihood. Acta Crystallogr D57:1373-1382, 2001.

89. B Rupp, BW Segelke. Questions about the structure of the botulinum neurotoxin B light chain in complex with a target peptide. Nature Struct Biol 8:643-664, 2001.

90. G Kleywegt, T Jones. Model Building and Refinement Practice. Meth Enzymol 277:208-230, 1997.

91. KA Kantardjieff, P Höchtl, BW Segelke, FM Tao, B Rupp. Concanavalin A in a dimeric crystal form: revisiting structural accuracy and molecular flexibility. Acta Crystallogr D58:735-43, 2002.

92. A Perrakis, TK Sixma, KS Wilson, VS Lamzin. wARP: Improvement and Extension of Crystallographic Phases by Weighted Averaging of Multiple-Refined Dummy Atomic Models. Acta Crystallogr D53:448-455, 1997.

93. CI Branden, TA Jones. Between objectivity and subjectivity. Nature 343:687-689, 1990.

94. RA Laskowski, MW MacArthur, DS Moss, JM Thornton. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26(2):283-291, 1993.

95. RRW Hoft, G Vriend, C Sander, EE Albola. Errors in protein structures. Nature 381:272-272, 1996.

96. GJ Kleywegt, TA Jones. Where freedom is given, liberties are taken. Structure 3:535-540, 1995.

97. M Paris. Vapornomics. Nature Biotechnology 19:301, 2001.

98. WP Burmeister. Structural changes in a cryo-cooled protein crystal owing to radiation damage. Acta Crystallogr D56:328-341, 2000.

99. M Weik, RBG Ravelli, G Kryger, S McSweeney, ML Raves, M Harel, P Gros, I Silman, J Kroon, JL Sussman. Specific chemical and structural damage to proteins produced by synchrotron radiation. Proc Natl Acad Sci USA 97:623-628, 2001.

100. VF Hartemann, HA Baldis, AK Kerman, LF A, NC Luhmann, Jr, B Rupp. Three-Dimensional Theory of Emittance in Compton Scattering and X-ray Protein Crystallography. Phys Rev E64:16501-1-16501-26, 2000.

101. R Henderson. Excitement over X-ray lasers is excessive. Nature 415:833, 2002.

102. Z Otwinowski, W Minor. Processing of X-ray Diffraction Data Collected in Oszillation Mode. Meth Enzymol 267:307-326, 1997.

103. J Westbrook, Z Feng, L Chen, H Yang, HM Berman. The Protein Data Bank and structural genomics. Nucleic Acids Research 31:489-91, 2003.

104. L Holm, C Sander. Protein Structure Comparison by Alignment of Distance Matrices. J Mol Biol 233:123-138, 1993.

105. PE Bourne, H Weissig. Structural Bioinformatics. Hoboken, NY: Wiley-Liss, 2003.

106. T Peat, E deLaFortelle, C Culpepper, J Newman. From information management to protein annotation: preparing protein structures for drug discovery. Acta Crystallogr D58:1968-1970, 2002.

107. B Rupp, BW Segelke, H Krupka, T Lekin, J Schafer, A Zemla, D Toppani, G Snell, T Earnest. The TB structural genomics consortium crystallization facility: towards automation from protein to electron density. Acta Crystallogr D58:1514-1518, 2002.

108. EA Merritt, DJ Bacon. Raster3D: Photorealistic molecular graphics. Meth Enzymol 277:505–524, 1997.

109. DE McRee. A visual protein crystallographic software system for X11/Xview. J Mol Graph 10:44–46, 1992.

110. JR Helliwell. Synchrotron radiation facilities. Nature Struct Biol 7:614-617, 1998.

111. VS Lamzin, A Perrakis. Current state of automated crystallographic data analysis. Nature Struct Biol 7:979-981, 2000.

112. CCP4. The CCP4 Suite: Programs for Protein Crystallography. Acta Crystallogr D50:760-763, 1994.

113. MD Winn, AW Ashton, PJ Briggs, CC Ballard, P Patel. Ongoing developments in CCP4 for high-throughput structure determination. Acta Crystallogr D58(11):1929-1936, 2002.

114. PD Adams, RW Grosse-Kunstleve, LW Hung, TR Ioerger, AJ McCoy, NW Moriarty, RJ Read, JC Sacchettini, NK Sauter, TC Terwilliger. PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr D58:1948-54, 2002.

115. AT Brünger, PD Adams, GM Clore, WL DeLano, P Gros, RW Grosse-Kunstleve, JS Jiang, J Kuszewski, M Nilges, NS Pannu, RJ Read, LM Rice, T Simonson, GL Warren. Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. Acta Crystallogr D54:905-921, 1998.

116. Edl Fortelle, G Bricogne. Maximum-Likelihood Heavy-Atom Parameter Refinement for Multiple Isomorphous Replacement and Multi-wavelength Anomalous Diffraction Methods. Meth Enzymol 276:472-494, 1997.

117. A Perrakis, R Morris, VS Lamzin. Automated protein model building combined with iterative structure refinement. Nature Struct Biol 6:458-463, 1999.

118. T Holton, TR Ioerger, JA Christopher, JC Sacchettini. Determining protein structure from electron density maps using pattern matching. Acta Crystallogr D56:722-724, 2000.

119. T Oldfield. X-LIGAND: an application for the automated addition of flexible ligands into electron density. Acta Crystallogr D57:696-705, 2001.

120. TA Jones, JY Zou, SW Cowan, M Kjeldgaard. Improved methods for the building of protein models in electron density maps and the location of errors in these models. Acta Crystallogr A47:110-119, 1991.

121. DE McRee. XtalView/Xfit - a versatile program for manipulating atomic coordinates and electron density. J Struct Biol 125:156-165, 1999.

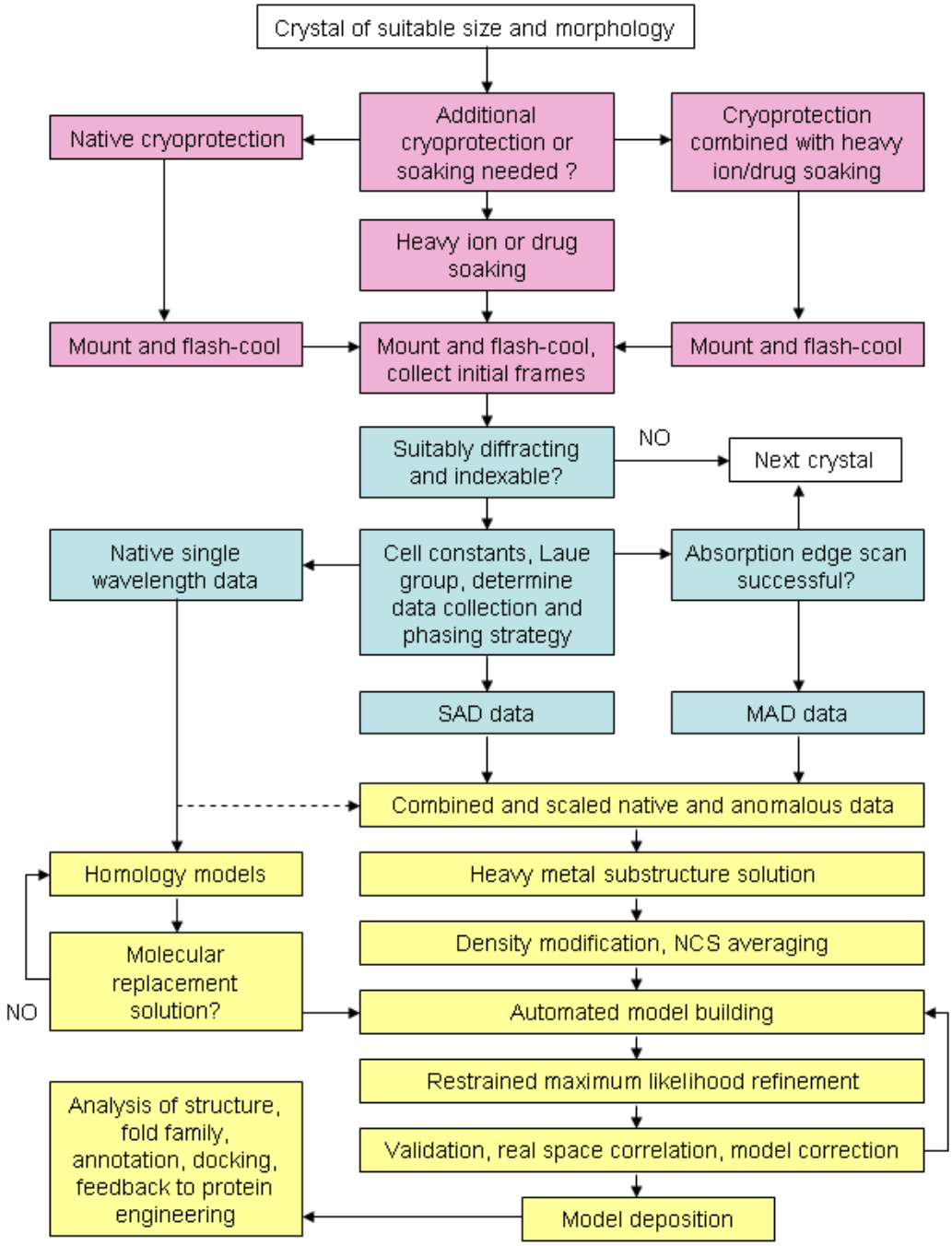
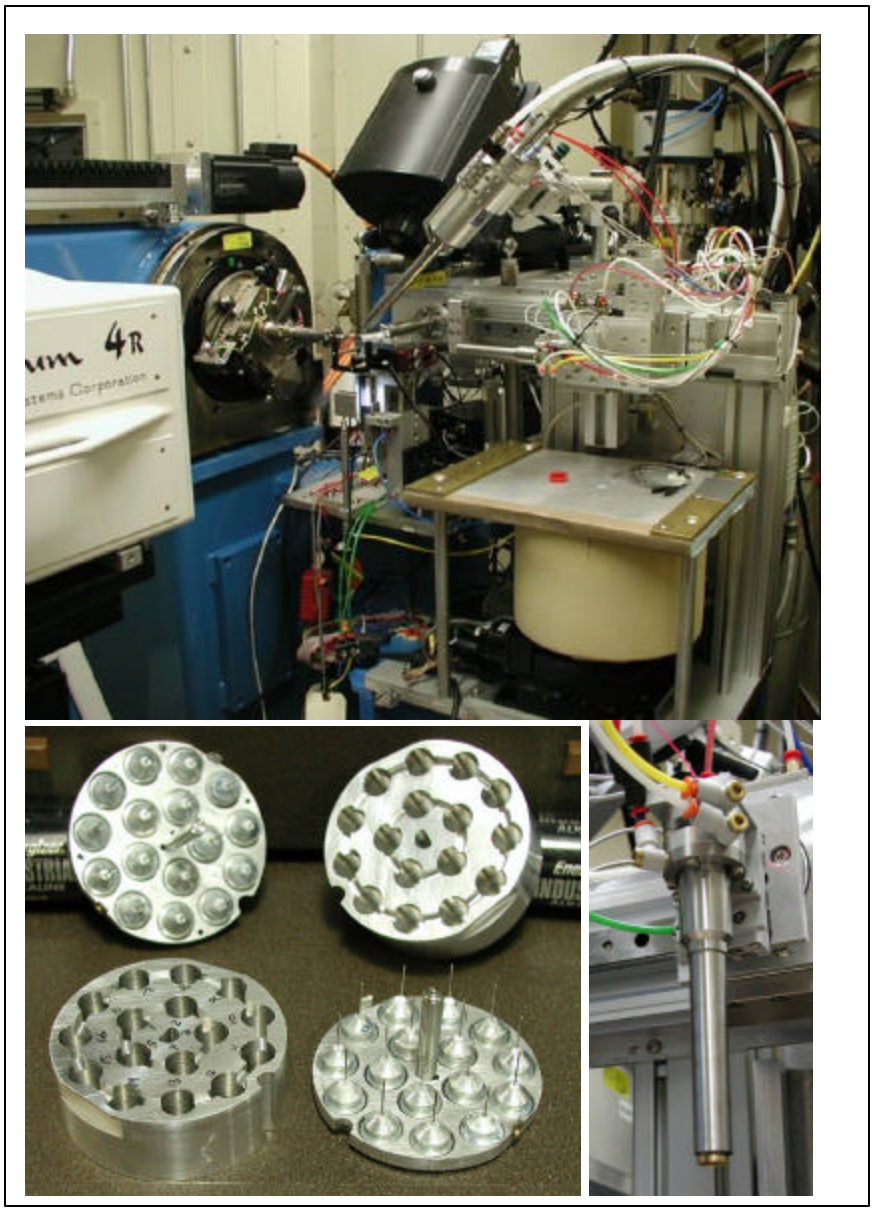


Figure 1: Flow diagram of the key steps in automated high throughput structure determination, from crystal selection to final structure model. The magenta colored boxes indicate steps involved in micromanipulation of crystals (section 5.2.2). Blue shade indicates experimental data collection steps conducted at the X-ray source (5.2.3). In yellow, steps that are conducted exclusively *in silico* (5.2.4).

Figure 2. **ALS developed automated sample mounting system**. Top: Overall view of sample mounting robot in hutch of beam line 5.0.3. at the Advanced Light Source (ALS) in Berkeley, CA, USA. Bottom left: detail view of pucks, 4 of each contained in the Dewar visible at the bottom of top panel. Bottom left: detail view of the pneumatically operated, cryo-cooled sample gripper, which retrieves the magnetic base sample pin from the Dewar and mounts them on the goniostat. The crystals in the sample loops are automatically centered on a motorized goniometer head (left side of instrument in top panel). Figure from (107) reproduced with permission from the IUCr.



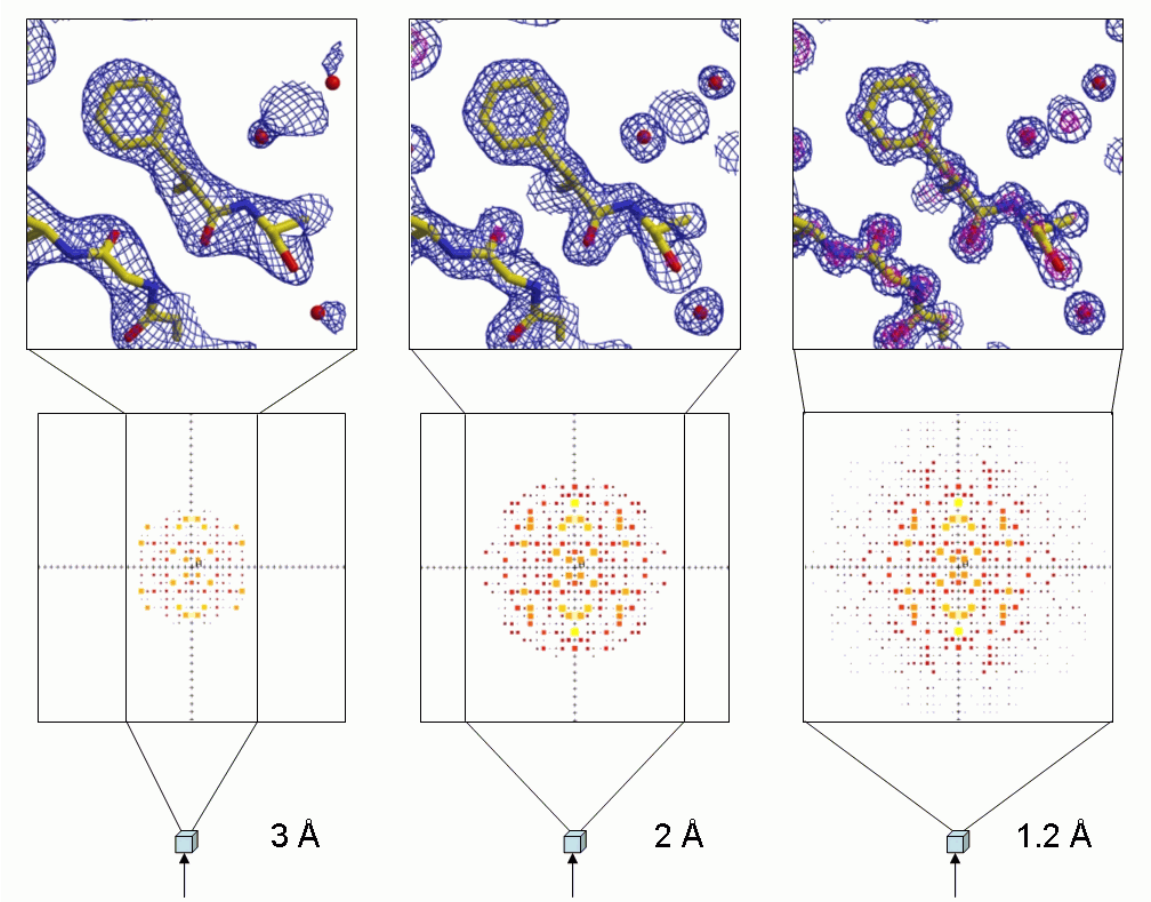


Figure 3: **Diffraction and electron density at increasing resolution.** The crystals (bottom) diffract X-rays increasingly better (diffraction limit or resolution of 3.0, 2.0 and 1.2 Å from left to right). Increasing diffraction (corresponding to finer sampling) produces many more reflections in the diffraction pattern (center row of figure) and hence, a more detailed reconstruction of the electron density map (blue contour grid) and building of a more accurate model is possible (top row). The modeled protein and ligand are shown in a ball and stick representation. Figures prepared using Raster3D (108) and XtalView (109).

Figure 4: **X-ray absorption edge**. An X-ray absorption edge scan for the anomalous marker atom is necessary for optimal wavelengths choice in MAD experiments. The theoretical position of the absorption edge (red saw tooth line) shifts due to varying chemical environment, and the scan is decorated with features stemming from electronic transitions in the X-ray Absorption Near Edge Spectrum (XANES) region and nearest neighbor scattering in the Extended X-ray Absorption Fine Structure (EXAFS) region. Absorption edges can have white lines, which result from electronic transitions into unoccupied atomic energy levels for elements with np or nd levels with n > 3. White lines contribute to a substantial increase in anomalous signal.

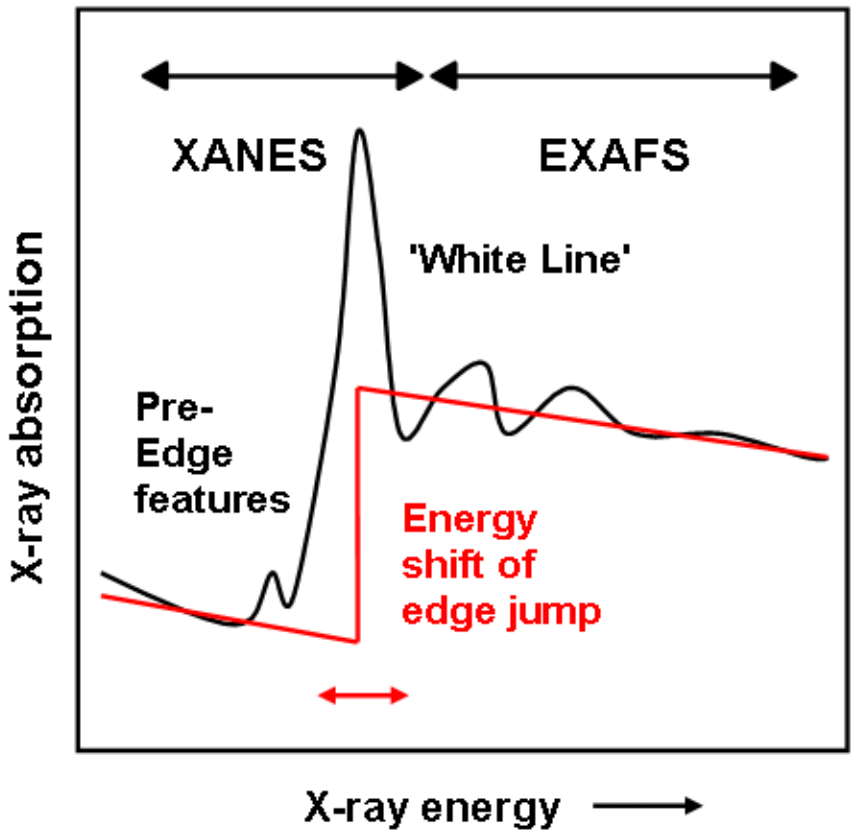
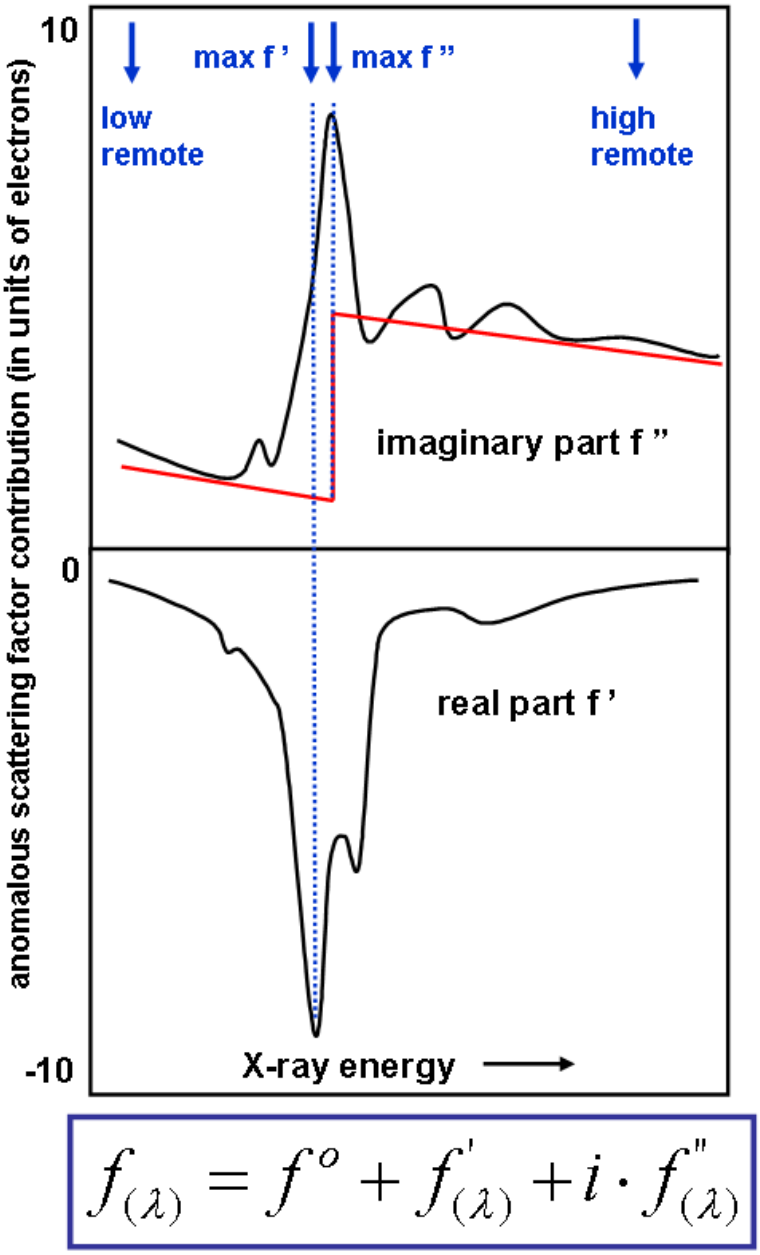


Figure 5: **Choice of MAD wavelengths**. The normalized X-ray absorption edge scan for the anomalous marker atom defines optimal wavelengths choice in MAD experiments. Top Panel shows the imaginary contribution f" to the atomic scattering factor of the marker atom (Formula in box). The Kramers-Kronig transform (lower panel) shows a minimum at the inflection point of the edge and is located at the maximum of the real contribution f'. Note that the values for max f' and f" are close together, an exact edge scan is therefore necessary for optimal MAD experiments.



$$f_{(\lambda)} = f^{o} + f^{'}_{(\lambda)} + i \cdot f^{''}_{(\lambda)}$$

Figure 6. **The nature of the crystallographic phase problem**. Reconstruction of electron density $?_{(x,y,z)}$ via Fourier transformation (formula) requires two values for each reflection: The structure factor amplitude $|F_{hkl}|$, which is proportional to the square root of the measured reflection intensity and readily available, and the phase angle a_{hkl}, which is unknown. Additional phasing experiments need to be carried out to obtain the missing phases. The need to determine phases by other means than direct measurement is referred to as the Crystallographic Phase Problem.

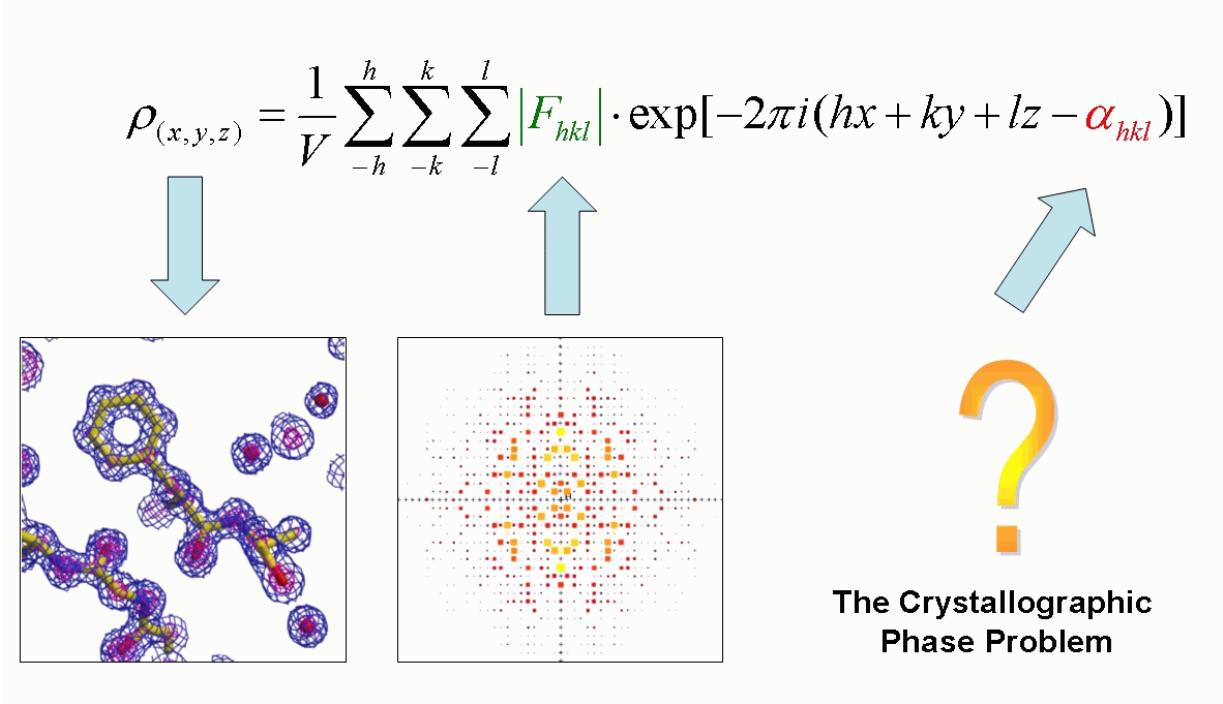


Figure 7. **The principle of crystallographic difference methods**. Top row presents the real space scenario, showing how differences simplify the search for an isomorphous marker substructure (big red atom). Left crystal, derivative, middle crystal, native, and right side, fictitious 'difference crystal'. The diffraction patterns (bottom row) are the reciprocal space representation of the real space scenario described above with crystals. In a similar way, anomalously scattering marker atoms create anomalous differences *within* a diffraction data set, and dispersive differences *between* data sets recorded at different wavelengths.

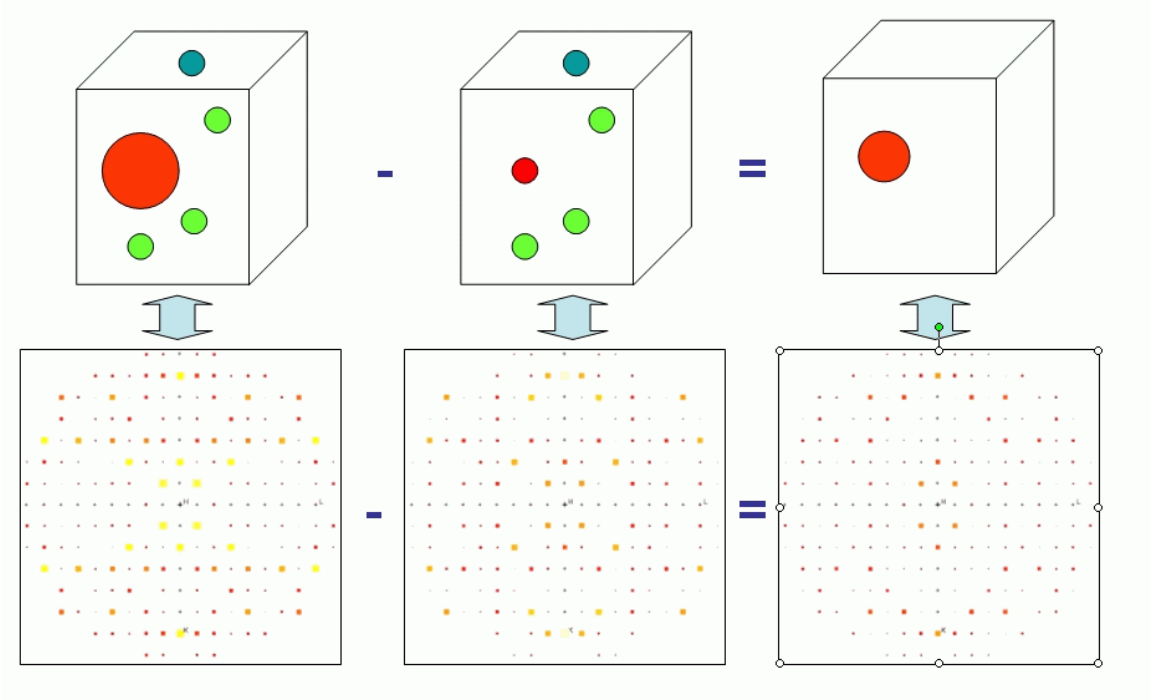


Figure 8. **Phasing via metal substructure.** Left panel: Harker section of a Patterson map created from the difference dataset in Fig.7. The marker atom positions are derived from distance vectors leading from the origin to the Patterson peaks. Right panel: Once the positions of the marker atoms are refined (represented by substructure vector F_H), the phase equations are solved and two possible solutions for the phase angle are obtained (circled). Degeneracy of the phase angle solution is resolved by using additional derivative structures, anomalous data, and/or density modification techniques.

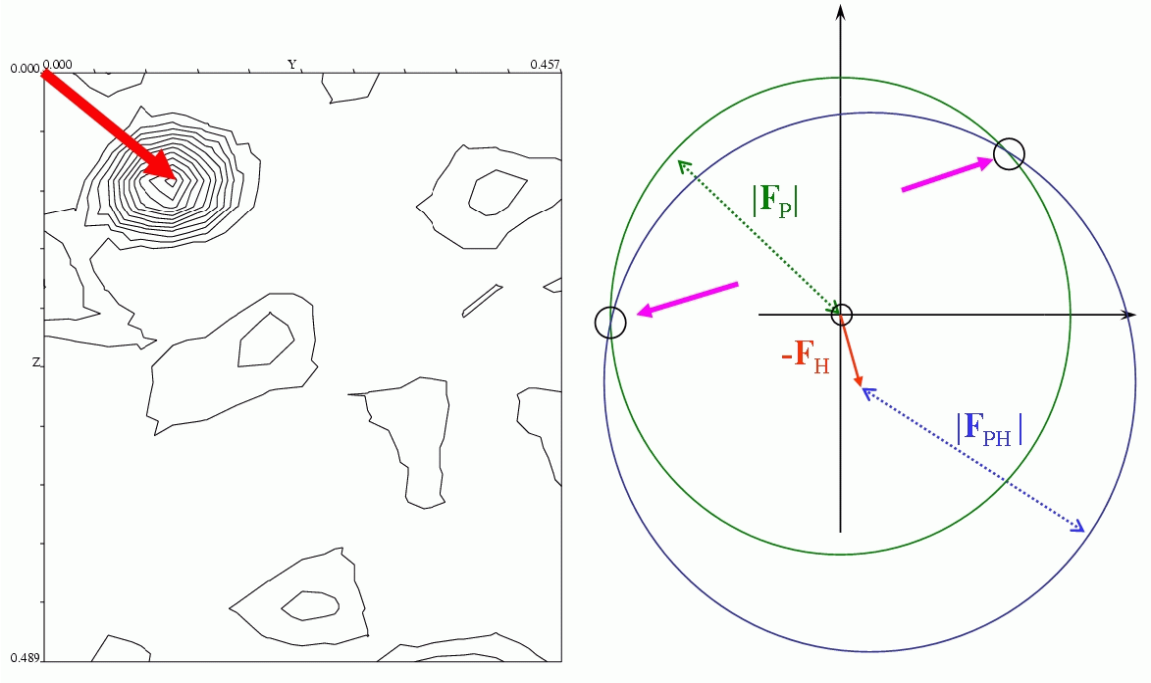


Figure 9. **Experimental electron density map after density modification.** The map shows a clear outline of the packed protein molecules, with solvent channels between them. Presence of these solvent channels allows soaking of small molecule ligands or drugs into protein crystals.

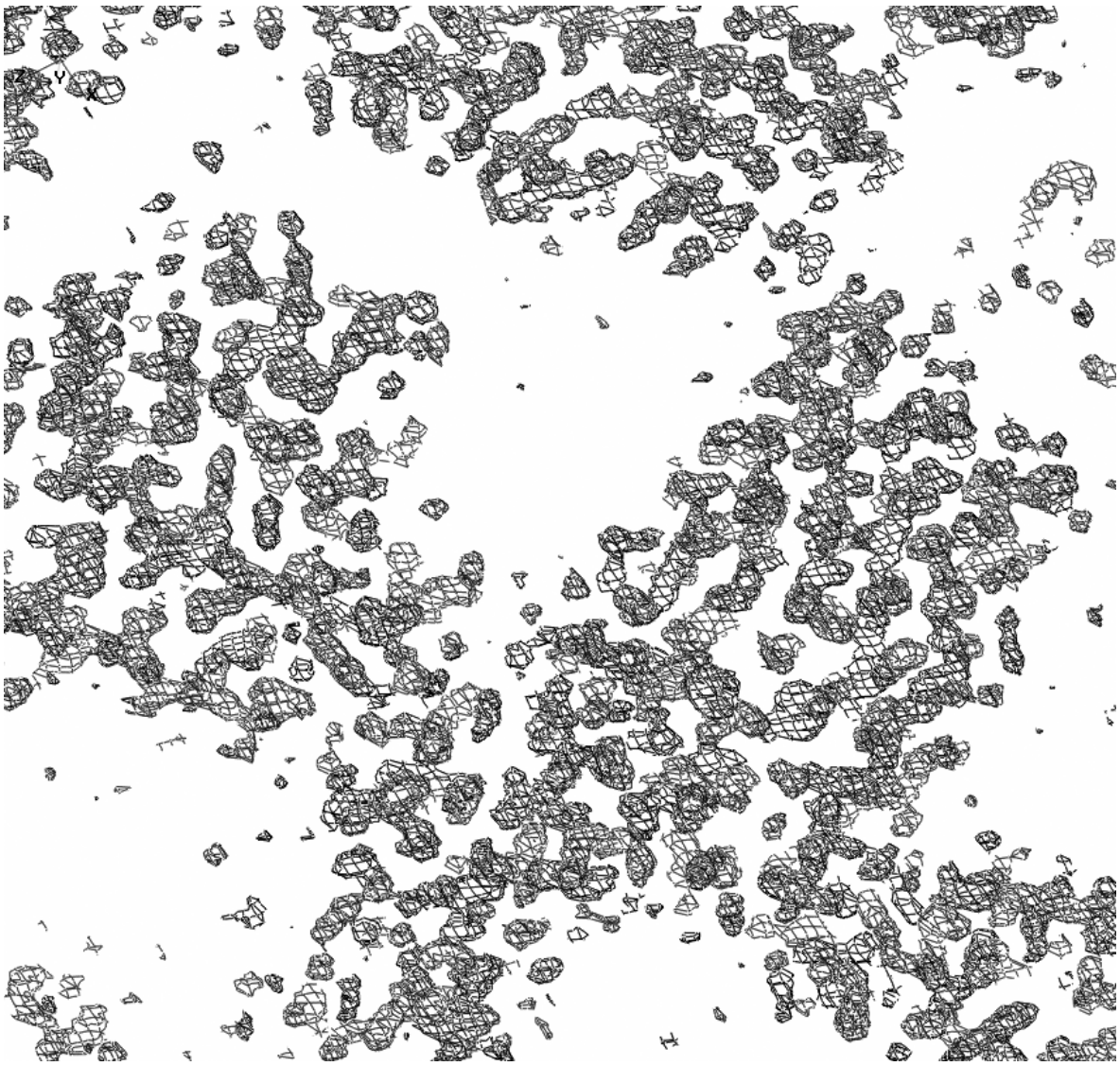


Table 1: Major Synchrotron facilities and beam lines equipped for HTPX. For a general listing of present and planned world wide synchrotron facilities, see (110). Beam line end stations are rapidly added or upgraded with robotics, thus an on-line search for new capabilities at the beam lines should to be conducted to augment this table. Good portals are http://www-ssrl.slac.stanford.edu/sr_sources.html and the Structural Biology Synchrotron Users Organization, http://biosync.sdsc.edu/, which provide periodically updated listings of all macromolecular crystallography beamlines. DOE: United States Department of Energy, NIH: National Institutes of Health.

Location	Link	Remarks	Microfocus/micro-collimated beamlines	HTPX beamlines w. robotics
ALS Advanced Light Source, Berkeley, CA, USA	bcsb.lbl.gov	DOE facility	-	5.0.1, 5.0.2 and 5.0.3 8.2.1 (ALS design (107), September 2003)
APS Advanced Photon Source, Argonne, IL, USA	www.aps.anl.gov/aps cars9.uchicago.edu/biocars www.sbc.anl.gov	DOE facility Excellent user facilities on site	19ID (40µm, ribosome 50S, 30S), 19BM (100µm), 14BM, 14ID	19BM, 19ID (December 2003), 14ID (planned for 2005) SGX 31 (Mar Robot)
SSRL/SPEAR Stanford Synchrotron	smb.slac.stanford.edu	DOE facility, DOE/NIH upgrade to SPEAR 3	50µm min beam size on all beam lines	1-5, 9-1, 9-2, 11-1, 11-3

Radiation Laboratory, Stanford, CA, USA		January 2004		(Jan 2004)
CHESS Cornel High Energy Synchrotron Source, Ithaca, NY, USA	www.macchess.cornell.edu	Non-DOE facility, NSF and NIH funding	F1 (special request only)	F1 (Sept 2003)
NSLS National Synchrotron Light Source, Brookhaven, NY, USA	www.px.nsls.bnl.gov	DOE facility	-	X12B (Sept 2003)
ESRF European Synchrotron Radiation Facility, Grenoble, France	www.esrf.fr/UsersAndScience/Experiments/MX/	Multi-national funding	BM14 (100µm) ID29 (40µm) ID13 (down to 5µm) ID23 (Fall 2003)	BM14 (EMBL design, Feb 2004) On all ID beam lines summer 2003-2004 BM30 (in-house)
SPring-8 Super Photon ring 8GeV	www.spring8.or.jp	JASRI and RIKKEN	BL41XU (100µm)	BL26B1, BL26B2

Hyogo, Japan				
DESY Deutsches Elektronen Synchrotron, Hamburg, FRG	www.embl-hamburg.de	EMBL outstation	- (2008 PetraIII upgrade)	X-13 (Mar Robot, 2004) BW-7B (in-house, EMBL design)

Table 2: Phasing methods in HTPX. SAD: Single-wavelength Anomalous Diffraction; MAD: Multi-wavelength Anomalous Diffraction; SIR: Single Isomorphous Replacement; MIR: Multiple Isomorphous Replacement; (AS): with Anomalous Scattering

Phasing Method	Phasing Marker	Derivatization Method	Remarks	Suitability for HTPX
SAD via sulfur atoms (S-SAD)	S in Met, Cys, residues, combined with solvent density modification	None, native protein	Requires highly redundant data collection	Becoming established, probably increasing use
MAD/SAD via naturally bound metals	Naturally bound metal ion, cofactor	None, native protein		Selected cases only
MAD via Se	Se in Se-Met residues	Incorporated during expression in met- cells or via metabolic starvation	1 Se phases 100-200 residues	Reliable standard, generally applicable, few exceptions
MAD via isomorphous metals	Heavy metal ion specifically bound	Soaking or co-crystallization	Hg, Pt, Au, etc Strong signal on L-edges due to XAS 'white lines'	Reliable, generally applicable
SIR(AS) via	Heavy metal ion	Soaking or co-	Phasing power	Reliable, nearly always in

isomorphous metals	specifically bound, density modification	crystallization	proportional to z back soaking necessary	combination with anomalous method
MIR(AS) via isomorphous metals	Heavy metal ion specifically bound	Soaking or co-crystallization	Multiple derivatives needed back soaking necessary	Reliable, often in combination with anomalous method
SIR(AS) via anions	Heavy anion specifically bound, Br-, I-, I3-	Mostly brief soaking, or co-crystallization	I derivatives also suitable for Cu source, possibly back soaking	Not fully established in HTPX, probably increasing use
SIR(AS) via noble gas	Noble gas specifically bound, Xe, Kr	Pressure apparatus	Xe XAS edge unsuitable for most MAD beam lines	Isolated cases so far
MR via model structure	None	None	Needs homology model with close coordinate r.m.s.d.	70% of cases, increasing. Subject to model bias, particularly at low resolution
Direct Methods	None	None	Atomic resolution, small size	Few cases, but important in metal substructure solution

Table 3: Computer programs and program packages commonly used in or developed for high throughput protein structure determination. HA: Heavy Atom; ML: Maximum Likelihood; MD: Molecular Dynamics; SA: Simulated Annealing; TLS : Torsion, Libration, Screw. Additional compilation of specific programs is available in (111).

Program	Reference, web site	Coverage	Remarks	Suitability for HTPX
CCP4 program suite	(112, 113) www.ccp4.ac.uk	Data collection, data processing, phasing, ML refinement, model building, validation. No MD refinement, but the only TLS refinement program	Current version 4.2.2, Graphical interface CCP4i. New release mid 2003. Multi-author collaboration.	Most common program package. No expert system. Can be scripted, many of its program modules found in and interface with other packages.
PHENIX	(114) www.phenix-online.org	In final version complete from data collection to model building, currently parts from phasing to model building	Author team includes XPLOR/CNS experts. Open Python source, industrial consortium members	Still under development, should be well suited for automation/expert system. Will include MD and simulated annealing refinement.
XPLOR/CNS	(115)	Program pioneering SA and MD in refinement of X-ray and NMR	HTML interface. CNX commercial version via	Academic development continued under

	cns.csb.yale.edu	data. Complete package including phasing and MR.	MSI/Accelrys	PHENIX project.
MOSFLM	(36) www.mrc-lmb.cam.ac.uk/harry/mosflm/	Data image processing, integration, reduction, and scaling via CCP4. Available on most beam lines.	Reliable indexing and data collection, freely available via CCP4	Developments under way to integrate expert system, fully automate data collection.
HKL2000/DENZO	(102) www.hkl-xray.com	Data collection, integration, reduction, scaling (SCALEPACK module). Commercial indexing/processing service available.	License fee also for academics. Interfaces with CCP4.	Next to MOSFLM most popular data collection software suite.
SOLVE/RESOLVE	(54) www.solve.lanl.gov	Combined ML HA solution, phasing, reciprocal space density modification, and model building program. Interfaces with CCP4.	Pseudo-SIRAS Patterson approach to substructure solution, ML HA refinement.	Easy to use and to automate via scripts. Interfaced with PHENIX.
SHARP, AUTOSHARP	(116) www.globalphasing.com	ML HA refinement, excellent phasing, density modification. Interfaces with CCP4.	Modern Server/Client architecture, Bayesian ML methods. Newer versions get faster.	Slower but more powerful phasing algorithm. Automated, links to ARP/wARP for

				model building.
ARP/wARP, warpNtrace	(117) www.embl-hamburg.de/ARP/	Map improvement via dummy atom refinement and model building. Works best at higher resolution.	Fully interfaced with CCP4/CCP4i. No source.	Fully automated model building, plans exist also for automated ligand building.
XPREP,SHELXD, SHELXE	(57) shelx.uni-ac.gwdg.de	HA data processing (XPREP), HA substructure solution by combined Patterson and direct methods, simpler but very fast phasing and density modification for maps.	SHELXD and SHELXE publicly available, XPREP, XM, XE commercial (Bruker XAS). No source.	Very reliable, fast, most often used as front-end for HA substructure solution for subsequent HA ML refinement and phasing programs. Well updated.
Shake and Bake	(56) www.hwi.buffalo.edu/SnB/	Direct methods full structure or HA substructure solution via reciprocal-direct scape cycling	Has also been used for complete small protein structure solution	Stand alone, interfaces via HA file to SOLVE/RESOLVE
TEXTAL	(118) textal.tamu.edu:12321	Low resolution automated model building based on pattern recognition	Better performance at low resolution than dummy atom based methods	Incorporated in PHENIX, actively developed, target resolution of as low as 3.5 Å

MAID	(73) www.msi.umn.edu/~levitt/	Relatively new model building program, evaluated in Badger, 2003	Combination of building techniques, real space torsion angle dynamics	No GUI, suitable for integration and automation. Works also at low resolution
QUANTA	(119) Accelrys www.accelrys.com/quanta/	Commercial descendent from Biosym/Xsight and MSI programs, now Accelrys.	Monolithic, expensive, includes ligand building XLIGAND	Well ntegrated package that delivers most of the functionality needed for structure determination.
AUTOSOLVE	Astex Pharmaceuticals www.astex-technology.co.uk/autosolve.html	Complete package including ligand placing and refinement	Not publicly available	Highly automated
ELVES	Holton J. ucxray.berkeley.edu/~jamesh/elves/	Clever UNIX scripting searching local installation for program components and combining them into a very basic expert system.	One of the first attempts towards expert systems.	Limited to the performance of the other available programs.
O	(120) alpha2.bmc.uu.se/~alwyn/o_related.html	Descendant of first generation of pioneering graphic modeling programs.	Many associated additional programs and utilities, partly included	GUI, scriptable, steeper learning curve than X-fit. Interfaces with CCP4.

			in CCP4	Well supported user group.
XtalView/Xfit	(121) www.sdsc.edu/CCMS/Packages/XTALVIEW/	Complete package for basic data processing, brute force HA location, phasing, and semi-automated model building (XFIT)	XFIT module allows easy model building and correction. Also Windows version WXFIT available	GUI, intended for fast manual building and correction of models. Easy to learn. Fast FFT support. Supported by CCP4i.
EPMR	(86) ftp.agouron.com/pub/epmr/	Evolutionary algorithm for molecular replacement, fast FFT allows 6d search.	Good convergence, automate search for multiple copies.	Well suited for multiple automated searches
Shake&wARP	(91) tuna.tamu.edu	Automated MR and bias removal program based on EPMR (Kissinger et al, 1997) and dummy atom placement/refinement via CCP4 programs.	Data preparation, MR, and multiple map averaging, bias minimized real space correlation. Source available.	Implemented as web service intended for automated structure validation. Slow but excellent, averaged bias minimized maps.