Acc. Chem. Res. **2003,** *36,* 173−181

High-Throughput Crystallography at an Affordable Cost: The TB Structural Genomics Consortium Crystallization Facility

BERNHARD RUPP*

Macromolecular Crystallography and TB Structural Genomics Consortium, Lawrence Livermore National Laboratory, Livermore, California 94551

Received May 7, 2002

ABSTRACT

The crystallization facility of the TB Structural Genomics Consortium, one of nine P50 structural genomics centers sponsored by the National Institutes of Health, provides TB consortium members with automated crystallization, data collection, and basic molecular replacement structure solution up to bias-minimized maps. In contrast to venture capital -funded commercial enterprises, the TB consortium facilities are decentralized and aim to develop high-throughput crystallography methods and automation on a comparatively small budget. In addition to financial constraints, the logistics and organization of a production environment differ considerably from academic settings. The TB Structural Genomics Consortium crystallization facility may thus provide a model for cost-effective, efficient high-throughput crystallography. Processes and methods presented in this review should assist academic institutions planning to invest in high-throughput structural biology to assess both the rewards and risks of establishing structural genomics programs.

1. Introduction

The TB Structural Genomics Consortium is a voluntary organization of researchers sharing a common interest in the structural biology of *Mycobacterium tuberculosis* (MTB), whose aim is to understand the structural basis for MTB pathogenicity.[1] In addition to individual efforts at various member laboratories, the TB consortium is supported by free access to National Institutes of Health−National Institute of General Medical Science-funded, decentralized core facilities[2] providing high-throughput cloning and protein purification [University of California, Los Angeles (UCLA), and Los Alamos National Laboratory (LANL)], crystallization [Lawrence Livermore National Laboratory (LLNL)], data collection [LLNL; beam line 5.0.2. at the Advanced Light Source (ALS) in Berkeley, and X8C at the NSLS in Brookhaven], and data warehousing (UCLA). Structures are solved at individual laboratories as well as at core facilities, depending on the arrangements with the consortium members who have targeted the particular gene. A main objective of the TB crystallization facility at LLNL is the development of affordable high-throughput crystallization techniques, with some emphasis on automated, homology-based molecular replacement structure solution techniques. The progress of the TB Structural Genomics Consortium is tracked through a public Web server (http://www.mbi.ucla.edu/TB). Consortium members receive automated weekly E-mail updates regarding the progress on their targets as well as statistics about the overall production of facilities and consortium member laboratories.

1.1. Structural Genomics in an Academic Environment. If one loosely defines structural genomics (SG) as the attempt to determine macromolecular structures on a genome-scale level, two orthogonal approaches are conceivable. One could acquire whatever resources are necessary to accomplish the selected task, or one might seek to optimize what could be achieved with the means provided. Given the potentially enormous rewards of structure-based drug development, it comes as no surprise that a substantial number of commercial ventures were able to attract the funds to approach the problem in the former way.[3] On the other hand, the recent public funding of SG efforts[2] provides, for the first time on a reasonable scale, the means for the development of nonproprietary high-throughput structure determination methods, which should benefit not only the funded centers but also, most importantly, any modestly sized academic structural biology effort.

Robotic automation is expensive, and full and complete automation quickly tends to become cost prohibitive—particularly in an academic environment. At the TB consortium crystallization facility, we therefore attempt to optimize the overall efficiency E of our process as defined in a simple linear model as

$$E = \frac{T \cdot S}{C}$$

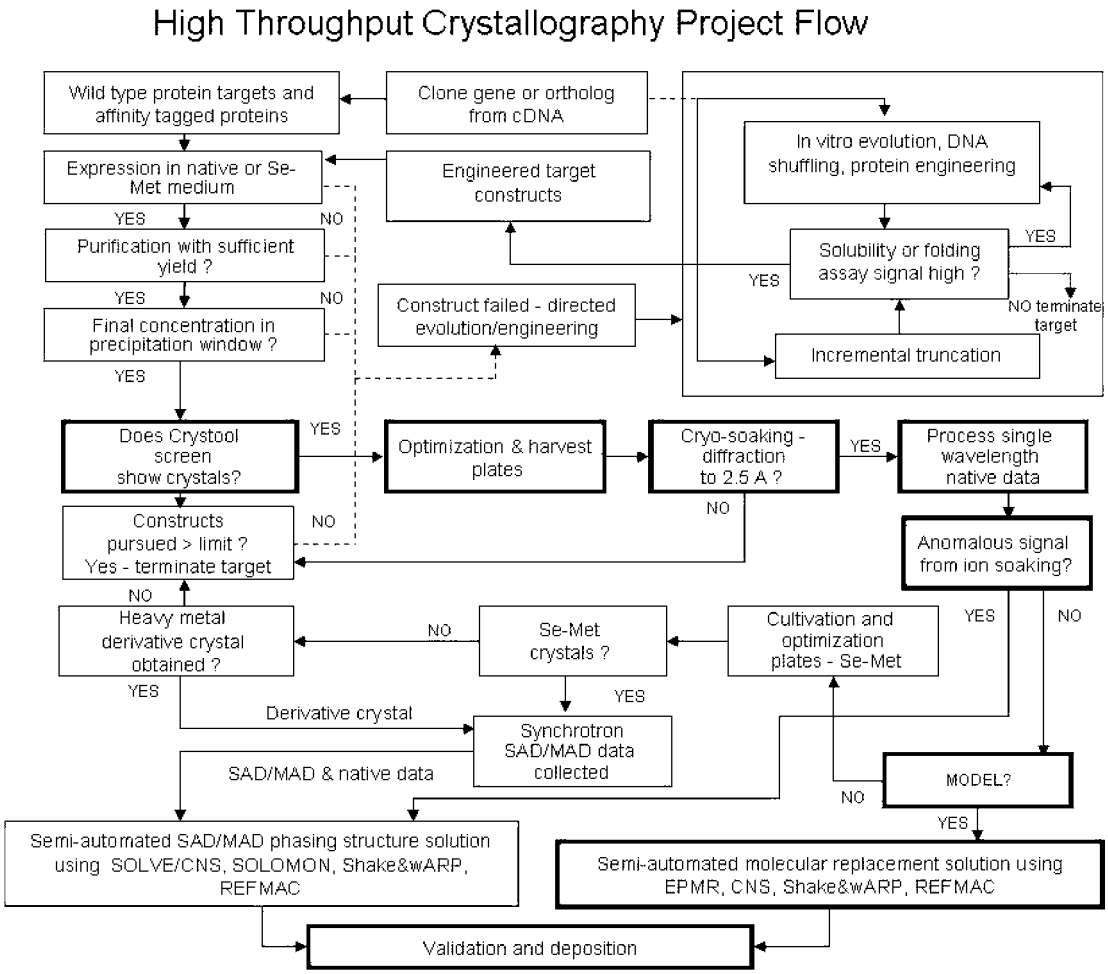
where T stands for throughput, S for success rate, and C for cost. Given cost as a (usually modest and limited) constant in a noncommercial environment, only T and S are viable candidates to increase E, the ultimate measure we chose as our academic (or NIH) shareholder value equivalent. The simple linear efficiency model, of course, cannot claim accuracy. Throughput itself may scale linearly with cost, and this low-risk approach has been taken frequently in academics by adding more graduate students to accomplish a proportionally higher throughput. In general, however, T and S are specific functions of C, and the efficiency model tends to become nonlinear. [Note: If the functions T and S are subject to feedback steps (e.g., quality control), the estimate of E, in fact, can

Bernhard Rupp, born January 1, 1956, in Vienna, Austria, graduated from the University of Vienna with a Ph.D. in physical chemistry in 1984, and received his venia docendi in molecular structural biology in 1998. After postdoctoral studies from 1984 to 1992 in structural chemistry and physics in Vienna, Zürich, Jerusalem, Jülich, and at the Lawrence Livermore National Laboratory (LLNL), he established the macromolecular crystallography facility at LLNL in 1993. Since 1997, he has been LLNL PRT research team leader at Advanced Light Source (ALS) beam line 5.0.2., where he conducted the inaugural MAD experiment of the ALS. In 2000, he formed the LLNL structural genomics group and established the TB consortium crystallization facility. Some of his interests are high-throughput crystallography, technology development, automated structure solution techniques, Compton X-ray sources, bacterial toxins and virulence factors, drug target structures, and machine learning. He also is a certified flight instructor and type rated airline pilot in the Boeing 737 and the Citation Jet.

* E-mail: br@llnl.gov.

10.1021/ar020021t CCC: $25.00
Published on Web 12/28/2002

Chart 1. Flow Diagram Relating Some of the Major Steps in High-Throughput Crystallography[a]



[a] Initial bio-informatics, feedback from analysis of structures, and process feedback from data mining, etc. are not included in this simplified view. Items in bold boxes represent tasks performed by the TB SG crystallization facility.

even become a coupled nonlinear system displaying divergent chaotic behavior. Interested readers may wish to consult an introduction to Operations Research for further references to linear programming as well as complex optimization and decision-making in economic models (for example, Hiller and Lieberman[39]).]

Operations in high-throughput production mode also require a skill (and mind) set distinctly different from that typical of classic, academic research. The wide range of disciplines involved in SG necessitates collaboration in diverse fields ranging from bioinformatics, protein engineering, robotics development, and applied mathematics to computer sciences. In particular, robotic engineering and data processing and tracking can grow into formidable tasks not easily handled in a "traditional" academic bioscience department, compounded with the scheduling and delivery demands of a high-throughput production environment. The NIH has clearly recognized these facts and noted that particular care must be taken when involving graduate students (and even postdoctoral fellows) in a high-throughput production and development environment (NIH−NIGMS RFA GM 99-009). In a publicly funded effort it is also, historically, difficult to argue that larger up-front investments (full-scale robotics, or information technology infrastructure improvements, for example) can provide long-term cost savings over the duration of the project or, at the same expense, allow more efficient use of the available resources (for example, graduate student talent can be engaged in intellectually more satisfying activity than high-throughput pipetting of crystallization cocktails).

Chart 1 provides a simplified overview of the process of structure determination in a high-throughput environment (note that the branches "terminate target" are frequently not an option for a hypothesis-driven academic environment, in particular if graduate students' careers are at stake). Nevertheless, it should become evident that there are many points where decisions must be made regarding which way to proceed. Points of decision harbor the great danger to become areas of nondecision, in which one tends to waste money in the hope of salvaging a project already exhibiting warning signs. In general, we can treat a SG project as sequences of steps, at which branching decisions determine how to proceed. Any branched decision tree can be reduced to a binary tree, and the formal treatment of a crystallographic structure determination process could be further refined by application of a Bayesian approach including priors modify-

ing the likelihood of outcomes.[4] Although full operations analysis has not yet been applied to SG, we will attempt to provide a rationale for each of the major decision points we faced in the development of the TB Structural Genomics Consortium crystallization facility.

In a process flow as shown in Chart 1, the total throughput is determined by certain rate-limiting steps, commonly described as bottlenecks. Due to progressing development of methods and techniques, rate-limiting steps do shift, and some extrapolation regarding the next expected rate limitations is sensible while tackling the current one. Although protein crystallization attracts much attention as a bottleneck today, the ultimate challenge of the future will be protein production, in particular once the "cherry-picking" phase of the SG initiatives evolves into targeting of membrane proteins, membrane-associated receptors, and larger, multiprotein complexes. At present, it appears fair to say that most proteins that are difficult to crystallize are also those that are hard to produce. Ordered phase separation, after all, requires solubility to begin with.

2. The TB Structural Genomics Consortium Crystallization Facility: Strategy and Implementation

2.1. Efficiency and Success Rate Analysis in Protein Crystallization. Crystallization is, in principle, a process of phase separation in a thermodynamically metastable system under the control of kinetic parameters. Given the unlimited number of combinations of components in crystallization recipes, it comes as no surprise that historically crystallization conditions were often chosen on the basis of what had worked before. Such screening kits based on previous success analysis have been quite popular,[5] and many variations of the first kit are now commercially available. In a statistical sense, repeated use of such premixed "sparse matrix" solutions amounts to oversampling of certain spots in a multidimensional crystallization space. Carter and Carter[6] very early recognized the need for a more rigorous statistical approach to crystallization screening and optimization by suggesting factorial designs and the use of variance analysis.[7] To optimally apply such rational methods, specific cocktails need to be prepared for each cycle of refinement, and unfortunately, when these advanced concepts were first introduced, use of robotics was not as widespread as it is now.

Segelke[8] has further assessed various crystallization screening protocols[5,9,10] in terms of sampling efficiency, i.e., finding crystallization conditions with a minimum number of trials. On the basis of rigorous statistical derivation, Segelke has shown that random (combinatorial) sampling is most efficient, particularly so when success rates are low or clustered. Efficiency analysis also allows estimating the number of trials above which return on investment (time, supplies, and protein) during further screening diminishes, as indicated by cumulative probability plots (ref 8, Figure 4). For the average soluble protein, as far as frequency and success rate data were available, we estimate that 288 (3×96) trials should suffice to find crystallization conditions with high probability. Past this point, the option of protein engineering (examples are discussed in refs 11−13) or search for orthologues should be investigated as a viable option, aiming to obtain an inherently more crystallizable variant[8] of the particular protein.

In random sampling, coverage of the crystallization space is achieved by using each crystallization condition only once. At the same time, prior knowledge about the specific protein and about success rate distributions can be included by customizing parameter ranges (pH, reagent concentrations) and frequencies. For example, with statistical evidence in favor of malonate as a precipitant,[14] the frequency of malonate in the combinatorial screen can be increased. In the same fashion, insight gained through rigorous statistical analysis of a sufficient number of experiments can be incorporated if the existence of "hot spots" in crystallization space is verified, while at the same time unsubstantiated or specifically unique "crystallization tips" will remain statistically insignificant noise (in more than one sense). One of the major scientific objectives of our crystallization facility is thus to create a comprehensive crystallization database through random sampling. Both the omission of negative result records and the lack of the most basic quantity in statistics, the number of trials, render the publicly available databases [Biological Macromolecule Crystallization Database[15] (BMCD), Protein Data Bank[16] (PDB)] virtually useless for the purpose of rigorous in-depth statistical analysis and machine learning. It must be understood, though, that even sophisticated statistical analysis and data mining of crystallization space (e.g., cluster analysis,[17] knowledge discovery, and case-based reasoning[18]) will only provide a basis for increasing the probability of crystallization success, but will not guarantee success for any particular protein.

As a consequence of de novo cocktail design for each protein construct, a large number of crystallization cocktails need to be prepared for screening and optimization. We thus implemented customizable random screen generation in the computer program CRYSTOOL[19] and interfaced it with a liquid-handling robot to automatically produce crystallization cocktails in a 96-well format. Details of the protocol implementation and robotic interfacing will be provided elsewhere and are summarized as follows.

2.2. Crystallization Cocktail Preparation. A set of 90 manually premixed stock solutions, divided into four groups—precipitants, buffers, additives, and detergents—are used to create random crystallization cocktails. User-selectable pH ranges and reagent frequencies allow inclusion of prior information when available. CRYSTOOL creates a set of procedure and performance files interpreted by the software of a Packard Instruments MPII-HT liquid-handling robot. Any liquid-handling station with independent, washable, stainless steel Teflon-coated variable-span tips with a useful dispensing range of 1 μL−1 mL can be used for this purpose. Liquid-level sensing and variable tip separation accommodate custom stock re-

agent racks (volumes of stock reagents required vary widely) as well as standard, Society for Biomolecular Screening (SBS)-compliant labware. Performance files account for varying liquid viscosities and associated wash and dispense requirements, and volatile components are dispensed last.

Production of de novo random screens is time-consuming (20−40 min per 96-well cocktail block) and is de facto rate limiting in our high-throughput crystallization process. To balance the desired comprehensive coverage of the crystallization space with the throughput requirements, we can use each of the 288 condition random screen sets for up to three different proteins. Such modest oversampling does not compromise the validity of random sampling data but allows us to conveniently screen about 10−20 protein samples per day, with the option of another 2-fold increase at a higher oversampling rate. Not unexpectedly, the true rate limitation for the near future appears to remain the availability of protein. Any excess capacity of our crystallization robotics, however, provides opportunity for systematic studies of factors such as temperature, setup technique and methods, batch variation, etc. In particular, small differences in success rate under different setup or environmental conditions require large sample sets to become statistically significant, and as a consequence, systematic studies of such effects are largely absent in the literature or limited to specific cases. In a limited preliminary experiment using five proteins with 96 random screens and different hanging and sitting drop setups, we could not establish significant differences in overall success rate, although individual conditions do show variation.

2.3. Protein Prescreening. Given an established correlation with crystallization success, inclusion of prior information about protein properties should enable an estimate of the likelihood of success of crystallization for a given protein under certain conditions. Light scattering (LS) allows determination of hydrodynamic properties of a protein, which can influence its propensity to crystallize, and LS has been proposed as a tool to predict crystallization success.[20] Although the usefulness of dynamic LS (DLS) techniques to study the physicochemical properties of a particular protein and of protein crystallization appears well established,[21] use of DLS as a rapid, high-throughput screening tool is probably limited, and we do not routinely employ it at the TB crystallization facility. The second virial coefficient (B_{22}) of a protein in a specific (crystallization) solution, determined in a series of static LS (SLS) experiments, can be interpreted as a measure of whether a protein falls into a "crystallization window".[20] Derived from the thermodynamic excess properties, B_{22} indicates the presence of attractive interactions between protein molecules in a given solute but cannot predict whether kinetics will actually favor crystallization. Similarly, even though monodispersity and narrow size distribution established by DLS do correlate with crystallization success rate,[21] and supporting atomistic explanations from AFM studies exist,[22] the percentage of proteins crystallizing from polydisperse or even impure solutions is still significant, probably around 30%. After all, protein crystallization was used as a purification technique long before the advent of protein crystallography (J. B. Sumner, 1919, see ref 23), and the cheapest, fastest, and by far most conclusive measure for crystallization still remains—crystallization screening.

We do, however, employ a quick test for solubility to establish whether a protein lies within a "precipitation window". The quest for the optimal concentration for a specific protein is still open, but it is reasonable to assume that when each representative of a precipitant class (ammonium sulfate as a salt, PEG 4000 as polymer, and ethanol as an organic solvent) at highest concentration fails to precipitate the protein, further concentration of the sample is advisable before full screening commences. Overall efficiency also decreases if effort is spent in concentrating a protein to some unsubstantiated magic value (10 mg/mL?) if precipitation prescreening indicates that 2 mg/mL might already suffice. We thus distribute a small prescreening kit, consisting of a 3×3 matrix of three precipitants in three increasing concentrations to consortium members and the production facilities. Similar tests have long been used in other laboratories (J. Jancarik, University of California, Berkeley; E. Cedergren-Zeppezauer, University of Lund, personal communications).

2.4. Crystallization Plate Setup. Requirements for dispensing precision, volume, and speed differ substantially for cocktail production compared to the actual plate setup. Fast, small (microliter to nanoliter)-volume, and very accurate (also in geometric terms) dispensing is mandatory for plate crystallization setup, whereas large-volume (milliliter) handling with modest speed and precision requirements suffices for cocktail setup. We thus decided, at the expense of full integration, to separate the plate setup from the cocktail mixing step. Once the cocktails are produced in a 96-well format, simple one-to-one dispensing into reservoir wells and drop aliquots into drop wells suffices. Various systems have been described which accomplish this task.[24] By augmenting a Hydra multichannel dispenser with a contact-less, single-channel Innovadyne dispensing unit, we can rapidly set up the cocktail reservoirs (200 μL) and, without re-arraying losses, dispense the protein (500 nL−1 μL) into drop wells filled with precipitant aliquots (Figure 1). The whole process of plate setup can be accomplished in less than 90 s, with sufficient time for wash steps after the plates have been sealed. Even with ample allowance of 10 min for washing and manual reloading, at least 16 proteins per 8-h shift can be screened in 288 experiments. Due to the rapid setup, drop sizes down to a total of 500 nL appear reasonably achievable using this technique without need for a humidity-controlled environment. The Hydra-Plus-One combination appears to be a fast and relatively inexpensive solution to protein crystallization setup, provided that premixed screens (true random or sparse matrix type) in 96-well format are available. Robotic loading of blocks and transfer of plates to a sealer can be readily accomplished with any SBS-standard-compliant plate crane if desired.

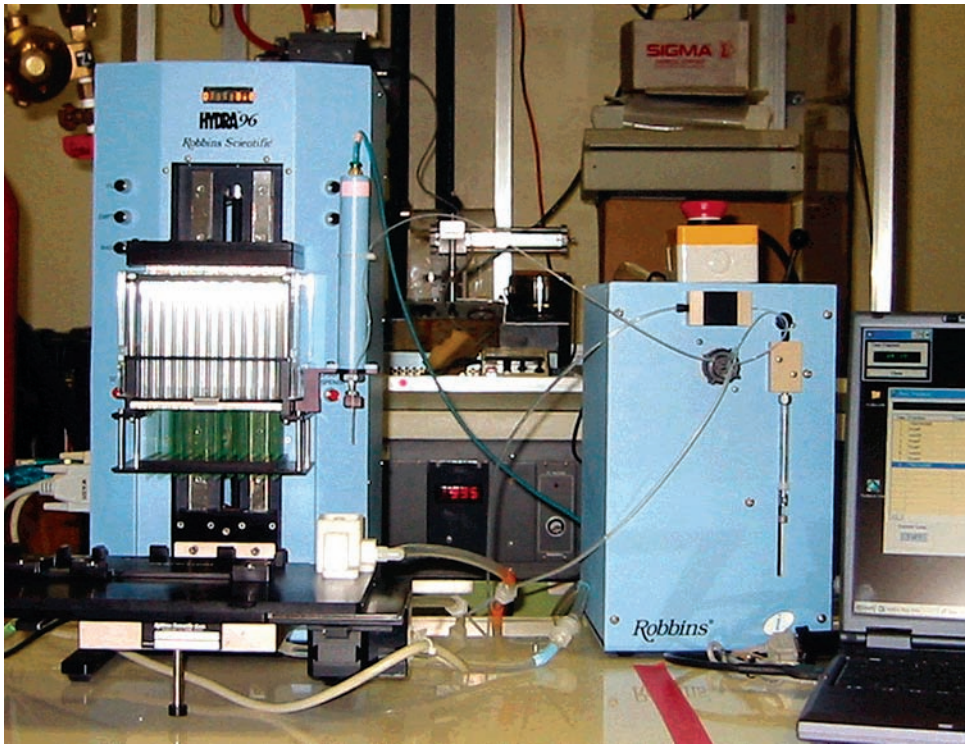


FIGURE 1. The Hydra-Plus-One crystallization robot. The integration of a standard 96-syringe (300 or 100 μL) Hydra liquid dispenser, an *xyz* table (blue unit on the left side of the picture), and a single-channel Innovadyne microsolenoid dispenser for the protein (blue control unit on the right, dispensing nozzle mounted on Hydra) allows rapid, contact-less dispensing of protein without the need for protein re-arraying.[42] Starting from premixed cocktails and aspirated protein, the total setup time per 96-well plate is 90 s.

2.5. Crystallization Plate Considerations. The choice of crystallization plate and method can be of substantial importance for the overall success of a high-throughput crystallization effort. We are not aware of any systematic studies demonstrating overall superiority of any one of the commonly used techniques (sitting drop, hanging drop, batch methods). [Note: Small droplet sizes, ease of miniaturization, and absence of additional sealing requirements should favor microbatch methods under oil for crystallization screening, as demonstrated by D'Arcy,[21] Luft et al.,[40] Chayen,[41] and others. Transfer to a different technique for optimization and harvesting of crystals, however, adds to the overall cost function in an integrated process. Similar considerations currently hold for another potentially promising microcrystallization development based on multilayer soft lithography (see www.fluidigm-.com).] For simplicity of handling and ease of harvesting, we elected to use the same sitting drop setup throughout for screening, optimization, and harvesting. We designed a suitable, SBS-compliant, 96-well plate for sitting drops, IntelliPlate, that specifically accommodates the needs of our high-throughput process. Details and results of a comparison with other plates will be presented elsewhere, but the main features can be summarized as follows. The plate has wide, elevated rims for reliable tape sealing and different well sizes to accommodate various drop sizes or additional cryobuffer during harvesting; polished wells prevent sticking of crystals and support easy harvesting; and well shape and optical properties are optimized toward automated image acquisition and recognition systems. Drop support and the drop itself form an optical system, and varying viscosities, surface tension, and wetting properties remain challenges for optimal (and universal) well design.

2.6. Image Acquisition and Crystal Detection Software. Based on a conservative throughput of 10 proteins screened per day, at 288 wells per protein (three 96-well plates) and a viewing schedule of seven times through the 6-month lifetime of a plate, we will accumulate plates up to a steady state in which an image of a crystallization experiment must be taken and analyzed approximately every 2 s during an 8-h shift. We thus consider image acquisition and analysis a high priority for full automation.

Our image acquisition hardware, VersaScan, configurable for any type of crystallization plate, is under development in collaboration with Velocity11 in Palo Alto, CA. Using the IntelliPlate, we can acquire a megapixel black-and-white image in about 0.5 s. Producing megabytes of data per second puts a definite strain on the data-processing systems, and reduction of raw data flow by intelligent analysis becomes a necessity. Significant progress has been made in several laboratories and commercial enterprises on crystal image analysis (for example, see ref 18). Based on multiple edge detection algorithms employing phase congruency and extensive pre- and postprocessing, we have developed a trainable system, with the ultimate objective of reliable crystal recognition, enabling subsequent automated optimization or harvesting plate

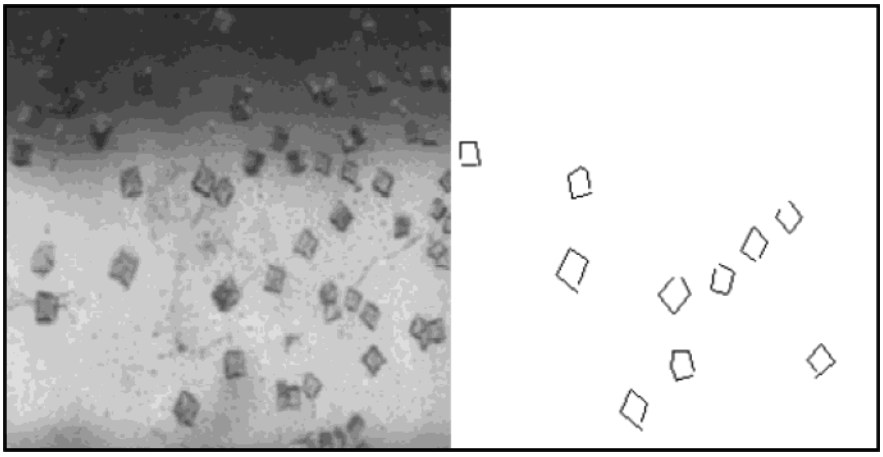


FIGURE 2. Crystals of TB protein Rv2523c. (Left) Raw image of small crystals in precipitate, acquired with automated VersaScan system, processed with new phase-congruency-based crystal recognition software (right). Despite low contrast, precipitate in background, and large magnification resulting in noisy, unfocused image, the recognition software detects sufficient features to identify a large number of small crystals, allowing reliable, automated scoring.

setup via CRYSTOOL without the need for human intervention (Figure 2).

Our basic handling unit for crystallization plates is one 48-plate rack, capable of accommodating the maximum achievable daily throughput of our system. A plate crane delivers the plates to the VersaScan image-processing unit and stacks observed plates into a second rack, which is manually returned to a temperature-controlled incubator. Large temperature differences between observation stage and plates give rise to condensation in the sealed wells, and reasonably stable temperature control in the observation room is a necessity.

2.7. Crystal Harvesting and Robotic Diffraction Screening. Crystal harvesting in suitable cryoloops with magnetic bases has become an inexpensive and reliable de facto standard in cryocrystallography.[25] In addition to cryoprotection, sweeps in cryobuffers allow at the same time introduction of heavy metals or anions such as bromide and iodide as phasing markers. In particular, due to the location of metal or iodine L edges (or even uranium M edges) not too far below the characteristic Cu Kα wavelength, in-house SAD/SIRAS phasing[26] should become an increasingly interesting alternative to synchrotron based multiwavelength methods.

Full automation of harvesting micromanipulations appears cost prohibitive at present for all but the most affluent industrial or large facility installations. [Note: A good example illustrating the high cost of automated mounting is the Crystal Preparation Prime Item (CPPI) developed by Oceaneering Space Systems for NASA (http://www.oceaneering.com/adtech/space/adtech_space_crystalprep.htm).] We currently are not attempting any automation of crystal harvesting in cryoloops (although an optimally designed crystallization plate reduces the effort spent in the process). Should crystals become so plentiful that harvesting develops into the rate-limiting step, the proven success at that point justifies further substantial investment in high-throughput robotic crystal harvesting. Novel microdevice support systems for nanodrops have recently been introduced,[27] and although challenges such as efficient cryocooling and protection remain, innovative approaches might eventually lead to automated, in situ diffraction screening techniques.

Fully automated mounting of the cryopins on the diffractometer, however, does greatly enhance utilization of valuable synchrotron (and laboratory source) beam time, and reliable commercial systems are becoming available (MAR Research, Rigaku/MSC). Under the assumption that any crystal deserves screening, fast and reliable storage and mounting procedures are necessary to realize high-throughput data collection for macromolecular crystallography. At the TB consortium crystallization facility, we use the sample transport and storage system developed at the ALS Macromolecular Crystallization Facility[28] on beam line 5.0.3. The basic handling unit, a cylindrical, puck-shaped cassette containing 16 standard cryopins, also serves as an integral part of a complete, automated cryogenic sample alignment and mounting system. Seven puck cassettes fit into a standard dry-shipping dewar. The mounting robot can select samples with a cooled robotic gripper from four pucks placed in a liquid nitrogen vessel. Mounting takes approximately 10 s, during which the crystal temperature is maintained below 110 K. Crystals are centered automatically through a centering algorithm on a remote-controlled goniometer head.

Our data collection strategies are conventional, with some allowance for lack of (still human) judgment in the early morning hours. Lower Laue symmetries are selected in any case of doubt about the space group, and a second, fast low-resolution sweep to avoid pixel saturation is becoming standard practice. The need for high-quality, low-resolution data for any phasing method (including molecular replacement, MR) has been pointed out repeatedly.[29] For the ease of model building and successful use of automated procedures, except in special cases, we do not collect or process MR data sets with resolution worse than 2.5 Å, but rather we pursue additional crystallization optimization. We estimate that our overall throughput is greater using the high-resolution strategy in view of the

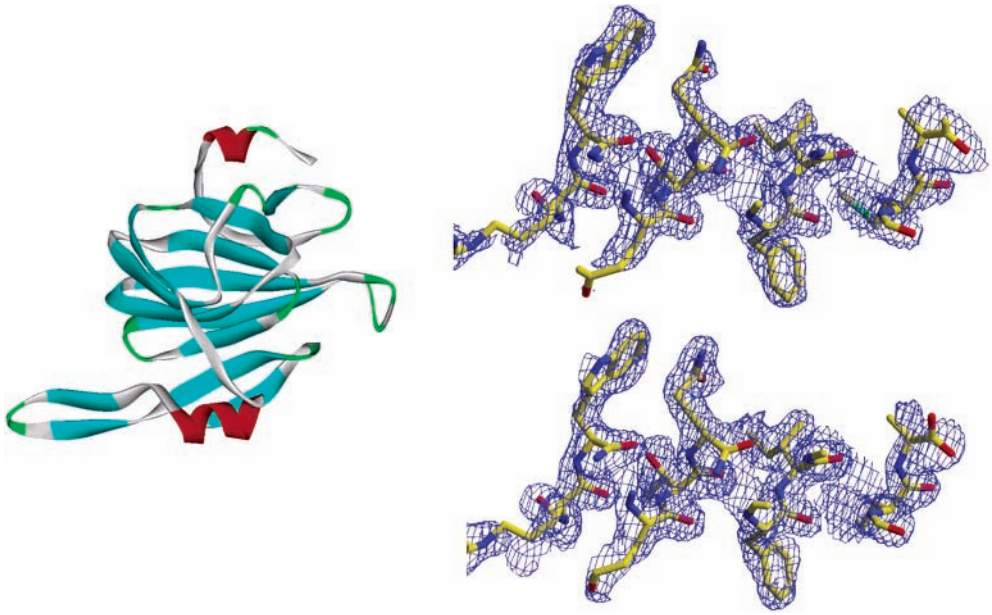


FIGURE 3. MTB Rv3465 protein structure. Rv3465, a dTDP-4-dehydrorhamnose 3,5-epimerase, was the first structure entirely processed by facilities of the TB Structural Genomics Consortium. (Left) Ribbon diagram of molecular structure. (Right) Initial electron density maps of unmodeled C-terminal region (upper small red helical region in ribbon diagram) after MR structure solution from a homology model. (Top right) Refmac5 $2mF_o - DF_c$ maximum likelihood map. (Bottom right) Bias-reduced Shake&wARP map. The final model (not used in map calculations) is superimposed on both maps. Improved side-chain definition and connectivity at the same level in the S&W map increase the ease and convergence of model building. Maps are contoured at one electron density level and displayed within 1.75 Å of the final model.

increased difficulty to accurately build and refine low-resolution models.[30]

2.8. Automated Molecular Replacement. Once native data are obtained either in-house from larger crystals or at the synchrotron, the availability of a homology model opens the possibility for MR phasing. Every successful MR solution might indeed save an additional phasing experiment. One needs to be aware, however, that much time can be wasted trying to rescue marginal MR solutions, only to arrive at a highly biased model that refuses to converge in refinement to a reasonable free *R*. The fact that a model delivers a weak MR solution does not yet mean a good structure will result quickly. It is equally important to subject the model, if necessary in repeated cycles, to effective bias removal techniques, as the effects of model (phase) bias can be insidious (particularly at low resolution) and are not always recognized by commonly used global structure quality descriptors such as *R* and R_{free} values.[31] Given the anticipated rise in coverage of structural folds available in the public database due to SG efforts, and given innovations in the method increasing the radius of convergence for powerful MR programs,[32,33] MR will very likely see constantly increasing use.

To quickly decide whether to proceed to ab inito phasing, we evaluate the potential for obtaining a reliable MR structure using a relatively simple automated protocol based on MR and homology structure prediction. A set of possible template structures is identified with multiple standard sequence alignment tools and retrieved automatically from the protein structure database. Homology backbone models are built from each of the template structures using the AL2TS 3-D model-building system.[34] Parallel MR searches for each of the highest-scoring models using the six-dimensional evolutionary search program EPMR[35] are branched to a computer cluster, and the models are evaluated according to their correlation to observed data. A recent review suggests that fold recognition models, although steadily increasing in quality,[36] still may not produce successful MR probes. While in conventional homology modeling experimental verification often is not available (or desired), the immediate feedback possible through evaluation of the model against experimental data 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 appear to benefit from such experimental restraints.

After side-chain building for the best MR solution, particularly marginal ones can be refined by simulated annealing torsion angle molecular dynamics[32] to bring them within the convergence radius of Shake&wARP (S&W). S&W is a highly effective bias removal and map reconstruction protocol, which is our derivative (briefly described in ref 23) of the original wARP procedure.[37] The fit of the model against the resulting S&W electron density is displayed in automatically generated real-space correlation plots, allowing for a rapid assessment of the local model structure quality. The first entirely facility-processed structure of the TB consortium, in fact, has been automatically solved from a modest MR solution with a correlation coefficient of 0.32 to a high-quality, bias-minimized electron density map (Figure 3). Automated model-building efforts are rapidly progressing in a number of academic laboratories[30,38] and commercial SG enterprises. [Note: Development is so rapid that only a Web search of sites can give an accurate view of the progress.

In no particular order, see Structural Genomix, www.stromix.com; Syrrx, www.syrrx.com; Astex Technologies, www.astex.com; or Accelrys, www.accelrys.com for some commercial developments.] We expect to implement automated MR service for TB consortium members on a Web server cluster.

3. Conclusions

We hope that emphasis on process analysis and on overall efficiency, as we attempt to implement in the TB consortium crystallization facility, will contribute to readily available and adaptable high-throughput crystallization procedures and instrumentation, demonstrating that high-throughput structure determination is feasible even for small workgroups—and at a reasonable cost.

My collaborators B. W. Segelke, H. I. Krupka, T. Lekin, J. Schafer, and D. Toppani have significantly contributed to the development of the TB crystallization facility. Tom Terwilliger, LANL, manages the substantial logistics as the director of the TB Structural Genomics Consortium. The cloning and protein production facilities under J. Perry, C. Goulding, D. Eisenberg (UCLA), M. Park, and G. Waldo (LANL) have supplied a steady flow of proteins. P. Malik and co-workers have developed the TB consortium Web site and database at UCLA. Members of Jim Sacchettini's group at Texas A&M University have provided me with numerous drug target complex structures as test cases for the automated MR and bias removal procedures. LLNL is operated by University of California for the U.S. DOE under contract W-7405-ENG-48. This work was funded by NIH P50 GM62410 (TB Structural Genomics) center grant.

References

(1) Goulding, C. W.; Apostol, M.; Anderson, D. H.; Gill, S. D.; Smith, C. V.; Yang, J. K.; Waldo, J. S.; Suh, S. W.; Chauhan, R.; Kale, A.; Bachhawat, A.; Mande, S. C.; Johnston, J. M.; Baker, E. N.; Arcus, V. L.; Leys, D.; McLean, K. J.; Munro, A. W.; Berendzen, J.; Park, M. S.; Eisenberg, D.; Sacchettini, J.; Alber, T.; Rupp, B.; Jacobs, W., Jr.; Terwilliger, T. C. The TB Structural Genomics Consortium: Providing a Structural Foundation for Drug Discovery. *Curr. Drug Targets: Infect. Disord.* **2002**, *2*, 121−141.

(2) Norvell, J. C.; Zapp-Machalek, A. Structural genomics programs at the US National Institute of General Medical Sciences. *Nat. Struct. Biol. Suppl.* **2000**, *7*, 931.

(3) Harris, T. The commercial use of structural genomics. *Drug Discovery Today* **2001**, *6*, 1148.

(4) Bricogne, G. Methodology of High Throughput Crystallography. *Book of Abstracts*, ICCBM9, Jena, March 23−28, 2002, Abstract O-A5.2.

(5) Jancarik, J.; Kim, S.-H. Sparse matrix sampling: A screening method for the crystallization of macromolecules. *J. Appl. Crystallogr.* **1991**, *24*, 409−411.

(6) Carter, C. W., Jr.; Carter, C. W. Protein crystallization using incomplete factorial experiments. *J. Biol. Chem.* **1979**, *254*, 12219−12226.

(7) Carter, C. W., Jr. Experimental design, quantitative analysis, and the cartography of crystal growth. In *Crystallization of Nucleic Acids and Proteins*, 2nd ed; Ducruix, A., Giege, R., Eds.; Oxford University Press: New York, 1999.

(8) Segelke, B. W. Efficiency analysis of sampling protocols used in protein crystallization screening. *J. Crystal Growth* **2001**, *232*, 553−562.

(9) McPherson, A. *Preparation and analysis of protein crystals*; Wiley: New York, 1982.

(10) Stura, E. A.; Nemerow, G. R.; Wilson, I. A. Strategies in the crystallization of glycoproteins and protein complexes. *J. Crystal Growth* **1992**, *122*, 273−285.

(11) Waldo, G. S.; Standish, B. M.; Berendzen, J.; Terwilliger, T. C. Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* **1999**, *17*, 691−695.

(12) Dale, G. E.; Kostrewa, D.; Gsell, B.; Stieger, M.; D'Arcy, A. Crystal engineering: deletion mutagenesis of the 24 kDa fragment of the DNA gyrase B subunit from Staphylococcus aureus. *Acta Crystallogr.* **1999**, *D55*, 1626−1629.

(13) Edwards, A. M.; Arrowsmith, C. H.; Christendat, D.; Dharamsi, A.; Friesen, J. D.; Greenblatt, J. F.; Vedadi, M. Protein production: feeding the crystallographers and NMR spectroscopists. *Nat. Struct. Biol. Suppl.* **2000**, *7*, 970−972.

(14) McPherson, A. A comparison of salts for the crystallization of macromolecules. *Protein Sci.* **2001**, *10*, 414−422.

(15) Gilliland, G. L.; Tung, M.; Blakeslee, D. M.; Ladner, J. The Biological Macromolecule Crystallization Database, Version 3.0: New Features, Data, and the NASA Archive for Protein Crystal Growth Data. *Acta Crystallogr.* **1994**, *D50*, 408−413.

(16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(17) Farr, R. G., Jr.; Perrymann, A. L.; Samudzi, C. T. Re-clustering the database for crystallization of macromolecules. *J. Crystal Growth* **1998**, *183*, 653−668.

(18) Jurisica, I.; Rogers, P.; Glasgow, J. I.; Fortier, S.; Luft, J. R.; Wolfley, J. R.; Bianca, M. A.; Weeks, D. R.; DeTitta, G. T. Intelligent decision support for protein crystal growth. *IBM Syst. J.* **2001**, *40* (2), 248−264.

(19) Segelke, B. W.; Rupp, B. Beyond the Sparse Matrix Screen: A Web Service for Randomly Generating Crystallization Experiments. *Am. Crystallogr. Assoc. Meeting Ser.* **1998**, *25*, 78.

(20) George, A.; Wilson, W. W. Predicting protein crystallization from a dilute solution property. *Acta Crystallogr.* **1994**, *D50*, 361−365.

(21) D'Arcy, A. Crystallizing Proteins—a Rational Approach? *Acta Crystallogr.* **1994**, *D50*, 469−475.

(22) McPherson, A.; Malkin, A. J.; Kuznetsov, Y. G.; Plomp, M. Atomic force microscopy applications in macromolecular crystallography. *Acta Crystallogr.* **2002**, *D57*, 1053−1060.

(23) Kantardjieff, K. A.; Höchtl, P.; Segelke, B. W.; Tao, F. M.; Rupp, B. Concanavalin A in a dimeric crystal form: revisiting structural accuracy and molecular flexibility. *Acta Crystallogr.* **2002**, *D58*, 735−43.

(24) Vilasenor, A.; Sha, M.; Thana, P.; Brauwner, M. Fast Drops: A High throughput Approach for setting up protein crystal screens. *Biotechniques* **2002**, *32*, 184−189.

(25) Rogers, D. W. Cryocrystallography techniques and devices. *International Tables For Crystallography*; IUCr, Kluwer Academic Publishing: Dordrecht, The Netherlands, 2001; Vol. F, pp 202−208.

(26) Matthews, B. W. Heavy atom location and phase determination with single wavelength diffraction data. *International Tables For Crystallography*; IUCr, Kluwer Academic Publishing: Dordrecht, The Netherlands, 2001; Vol. F, pp 293−298.

(27) Sanjo, A.; Cacheu, R. E. New Microfabricated Device Technologies for High Throughput and High Quality Protein Crystallization. *Book of Abstracts*; ICCBM9, Jena, March 23−28, 2002, Abstract O-I.2.

(28) Snell, G.; Meigs, G.; Cork, C.; Nordmeyer, R.; Cornell, E.; Yegian, D.; Jaklevic, J.; Jin, J.; Earnest, T. Automatic sample mounting and alignment system for macromolecular crystallography at the Advanced Light Source. *J. Synchrotron Radiat.* **2002**, in press.

(29) Dauter, Z.; Wilson, K. S. Principles of monochromatic data collection. *International Tables For Crystallography*; IUCr, Kluwer Academic Publishing: Dordrecht, The Netherlands, 2001; Vol. F, pp 177−195.

(30) Perrakis, A.; Harkiolaki, M.; Wilson, K. S.; Lamzin, V. S. ARP/wARP and molecular replacement. *Acta Crystallogr.* **2001**, *D57*, 1445−1450.

(31) Kleywegt, G. J.; Jones, T. A. Homo crystallographicus-quo vadis? *Structure* **2002**, *10* (4), 465−472.

(32) Adams, P. D.; Panu, N. S.; Read, R. J.; Brünger, A. T. Extending the limits of molecular replacement through combined simulated annealing and maximum-likelihood refinement. *Acta Crystallogr.* **1999**, *D55*, 181−190.

(33) Read, R. J. Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr.* **2001**, *D57*, 1373−1382.

(34) Zemla, A. Automated 3D protein structure predictions based on sensitive identification of sequence homology, in preparation, http://predictioncenter.llnl.gov, 2002.

(35) Kissinger, C. R.; Gelhaar, D. K.; Fogel, D. B. Molecular replacement by evolutionary search. *Acta Crystallogr.* **1999**, *D55*, 484−491.

(36) Jones, D. T. Evaluating the potential of using fold-recognition models for molecular replacement. *Acta Crystallogr.* **2001**, *D57*, 1428−1434.

(37) Perrakis, A.; Sixma, T. K.; Wilson, K. S.; Lamzin, V. S. wARP: Improvement and Extension of Crystallographic Phases by Weighted Averaging of Multiple-Refined Dummy Atomic Models. *Acta Crystallogr.* **1997**, *D53*, 448−455.

(38) Holton, T.; Ioerger, T. T.; Christopher, J. A.; Sacchettini, J. C. Determining protein structure from electron-density maps using pattern matching. *Acta Crystallogr.* **2000**, *D56*, 722−734.
(39) Hillier, F. S.; Lieberman, G. J. *Introduction to Operations Research*, 7th ed; McGraw-Hill: New York, 2000.
(40) Luft, J. R.; Wolfley, J.; Jurisica, I.; Glasgow, J.; Fortier S.; DeTitta, G. Macromolecular Crystallization in a High Throughput Laboratory−the Search Phase. *J. Crystal Growth* **2001**, *232*, 591−595.
(41) Chayen, N. E. Comparative studies of protein crystallization by vapour-diffusion and microbatch techniques. *Acta Crystallogr.* **1998**, *D54*, 8−15.
(42) Krupka, H. I.; Rupp, B.; Segelke, B. W.; Lekin, T. P.; Wright, D.; Wu, H.-C.; Tood, P.; Azarani, A. The high-speed Hydra-Plus-One system for automated high-throughput protein crystallography. *Acta Crystallogr.* **2002**, *D58*, 1523−1526.

AR020021T