The TB structural genomics consortium crystallization facility: towards automation from protein to electron density

Bernhard Rupp,[a*] Brent W. Segelke,[a] Heike I. Krupka,[a] Tim P. Lekin,[a] Johana Schäfer,[a] Adam Zemla,[a] Dominique Toppani,[a] Gyorgy Snell[b] and Thomas Earnest[b]

[a] Macromolecular Crystallography and TB Structural Genomics Consortium, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA, and [b]Berkeley Center for Structural Biology, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. E-mail: br@llnl.gov

The crystallization facility of the TB (Tuberculosis) structural genomics consortium, one of nine NIH sponsored P50 structural genomic centres, provides TB consortium members with automated crystallization, data collection and basic molecular replacement (MR) structure solution up to bias minimized electron density maps. Crystallization setup of up to ten proteins per day follows the CRYSTOOL combinatorial screen protocol using a modular and affordable robotic design with an open architecture. Components include screen preparation, plate setup, automated image acquisition and analysis, and optimisation design. A new 96 well crystallization plate has been designed for optimal robotic handling while maintaining ease of manual crystal harvesting. Robotic crystal mounting, screening, and data collection are conducted in-house and at the Advanced Light Source (ALS) in Berkeley. A simple automated protocol based on MR and homology based structure prediction automatically solves modestly difficult problems. Multiple search models are evaluated in parallel MR and the best multi-segment rigid body refined MR solution is subjected to simulated annealing torsion angle molecular dynamics using CNS, bringing even marginal MR solutions within the convergence radius of the subsequent highly effective bias removal and map reconstruction protocol, Shake&wARP, used to generate electron density for initial rebuilding. Real space correlation plots allow rapid assessment of local structure quality. Modular design of robotics and automated scripts using publicly available programs for structure solution allow for efficient high throughput crystallography - at a reasonable cost.

Keywords: high throughput crystallization; structural genomics; crystallization screening; molecular replacement; phase bias removal

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, with the aim of understanding the structural basis for MTB pathogenicity (Goulding *et al.*, 2002). In addition to individual efforts at various member labs, the consortium is supported by free access to NIH-NIGMS funded, decentralized consortium core facilities (Norvell *et al.*, 2000) for high throughput cloning and protein purification (UCLA and LANL), crystallization (LLNL), and data collection (LLNL, ALS, BNL). Structures are solved at individual labs as well as by core facilities, depending on the arrangements with the consortium members having targeted the particular gene. A main objective of the TB crystallization facility at LLNL is the development of affordable high throughput crystallization techniques and of automated structure solution methods, in particular homology-based MR techniques. Full and complete robotic automation quickly tends to become cost prohibitive - at least in an academic environment - and we attempt to optimise the total efficiency E of our process, 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 non-commercial environment, only T and S are viable candidates to increase E, the ultimate measure we chose as our academic (or NIH) share holder value equivalent. A more detailed account of overall efficiency considerations in structural genomics efforts will be provided elsewhere (Rupp, 2002).

2. The TB Structural Genomics Consortium crystallization facility: strategy and implementation

2.1. Efficiency and success rate analysis in crystallization

Segelke (2001) has assessed various crystallization screening protocols in terms of sampling efficiency, i.e. finding crystallization conditions with a minimum number of trials. Based on a rigorous statistical derivation, Segelke has shown that of the compared protocols, random (combinatorial) screening is most efficient, particularly so when success rates are low. 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. For the average, soluble, protein, as far as frequency and success rate data are available, we have estimated that 288 (3x96) trials should suffice to find crystallization conditions with high probability. Past this point, the option of protein engineering (Waldo *et al*., 1999, for example) or search for orthologs should be investigated as a viable option, aiming to obtain an inherently more crystallisable variant 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. Consequently, a great number of crystallization cocktails need to be prepared *de novo*. We thus implemented customisable random screen generation in the computer program CRYSTOOL (Rupp & Segelke, 1998) and interfaced it with a Packard Instruments MPII liquid handling robot to automatically produce crystallization cocktails in 96 well format (Marsh BioBlocks, 1.5 ml wells). Details of the protocol implementation and robotic interfacing are provided elsewhere (Krupka *et al*., 2002a) and are summarized as follows.

2.2. Crystallization cocktail preparation

A set of 90 manually prepared stock solutions, divided into 4 groups: precipitant, buffer, additive, and detergent is used to create random crystallization cocktails with pH ranges and reagent frequencies selectable by the user, thus enabling inclusion of prior information when available. CRYSTOOL creates a set of procedure and performance text files which are interpreted by the WINPREP software of a Packard Instruments Multiprobe MP-II HT liquid handling robot. The MP-II has 8 independent, washable, stainless steel and Teflon coated variable span tips with a useful dispensing range of 1µl to 1 ml. Liquid-level-sensing technology and variable tip separation allow to accommodate both custom stock reagent racks (volumes of stock reagents required vary widely) as well as standard, SBS (Society for Biomolecular Screening) compliant labware. Varying liquid viscosities and corresponding wash and dispense requirements are considered in performance files of the WINPREP instructions, volatile components are dispensed last. The cocktails are prepared in 96 well format (Marsh BioBlocks, 1.5 ml wells), heat or pressure sealed, and mixed by inverting and shaking the sealed plates before storage.

Production of *de novo* random screens is time consuming (20-40 min per 96-well cocktail block), and de-facto time-limiting in our high throughput crystallization process. To balance the desired comprehensive coverage of the crystallization space with the throughput requirements, we use each of the 288 condition random screen sets for 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 two-fold increase at a higher oversampling rate. Not unexpected, the true rate limitation for the near future appears to remain the availability of protein.

The option to move supplies and finished crystallization cocktail blocks to and from the MP-II with a plate crane has been considered for implementation at the TB crystallization facility, but presently the need for automated plate manipulation is much more critical in the image acquisition and analysis stage described later.

2.3. Crystallization plate setup

Dispensing precision, volume, and speed requirements differ substantially for the cocktail production compared to the actual plate setup. Fast, small µl to nl-volume and very accurate (also in geometric terms) dispensing is mandatory for plate crystallization setup, whereas large, ml volume handling with modest requirements of speed and precision suffices for cocktail setup. We thus decided, at the expense of full integration, to separate the pate setup from the cocktail mixing step. Once the cocktails are produced in a 96-well format, simple one-to-one dispensing into 200 µl reservoir wells and 1 µl drop aliquots in drop wells suffices. The true and proven Hydra multi-channel dispenser performs this task reliably, and by augmenting it with a contact-less, single channel Innovadyne dispensing unit, we can rapidly and without re-arraying losses dispense the protein into the already cocktail filled drop wells (Figure 1). The whole process of plate setup can be accomplished in less than 90 seconds, with sufficient time for wash steps after the plates have been sealed. Even with ample allowance for ten minutes of washing and reloading, at least 16 proteins per 8 hr 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 humidity controlled environment. The Hydra-Innovadyne 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 (Krupka et al., 2002b). Plates and blocks can readily be loaded, and finished plates automatically transferred to a sealer, with any SBS-standard compliant plate crane if desired.

2.4. Crystallization plate considerations

The choice of crystallization plate can be of substantial importance for the overall success of a high throughput crystallization effort. We have designed a new, 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 are to be published elsewhere (Krupka *et al*., 2002c), but the main features can be summarized as follows: The plate has wide, elevated rims for reliable sealing, different well sizes to accommodate various drop sizes or additional cryo-buffer during harvesting; polished round wells support easy harvesting, and well shape and optical properties are optimised towards automated image acquisition and recognition systems.

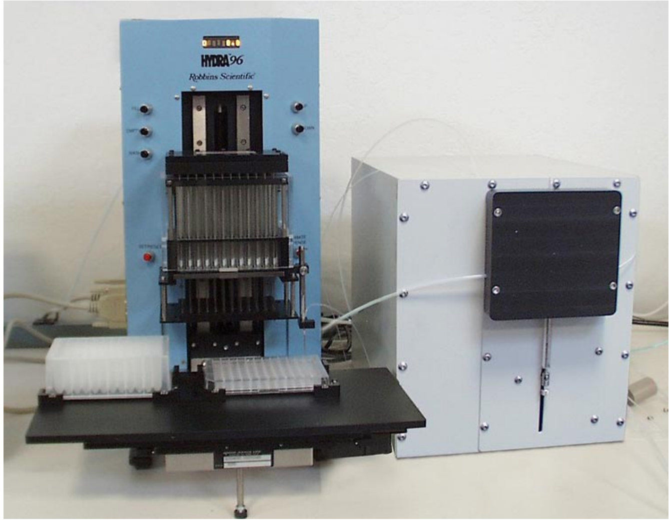

Figure 1

The Hydra/Innovadyne crystallization robot prototype in use at the TB consortium crystallization facility. A standard 96-syringe (300 or 100 µl) Hydra robot with xyz-table is simply combined with a single channel Innovadyne dispenser (right, white unit) allowing rapid, contact-less dispensing of protein without need for protein re-arraying. Total plate setup time, starting from premixed cocktails and aspirated protein, is 90 seconds.

2.5. 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 six months lifetime of a plate, we accumulate plates up to steady state in which an image of a crystallization experiment must be taken and analysed approximately every 2 seconds during an 8 hr shift.. We thus consider image acquisition and analysis as a high priority for full automation, including plate handling.

Our image acquisition system is a development prototype, VersaScan, designed in collaboration with Velocity11 in Palo Alto, CA. Using the IntelliPlate (the system is user configurable for any type of plate) we can acquire one mega pixel black and white image in about 0.5 seconds. The capability of producing two MB of (uncompressed) data per second puts a certain strain on the data processing systems, and reduction of raw data flow by intelligent analysis becomes a necessity. Progress has been reported in crystal image analysis (for example, Luft *et al*., 2001) and we are developing a proprietary, trainable system, with the ultimate objective that reliable crystal recognition as well as subsequent automated optimisation or harvest screens are set up without need for human intervention (Figure 2).

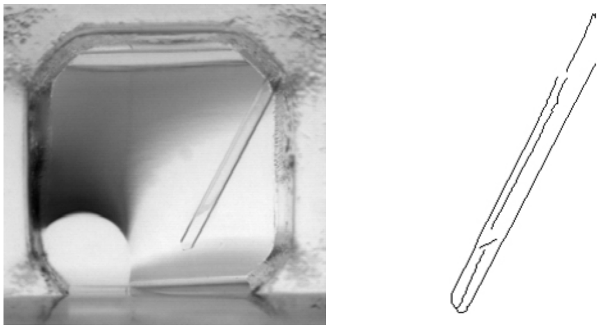

Figure 2

Crystals of TB protein Rv0773. Left panel, raw image of 4 µl drop in Greiner plate, acquired with automated VersaScan system. Note the low contrast as well as optical artefacts and suboptimal well format, necessitating further improvements in both acquisition hardware and plate design. Right panel, features extracted from raw image by crystal recognition software. Despite low contrast data, features are clear and allow reliable scoring of crystal properties. Progress in image analysis in several laboratories will likely deliver stable recognition and scoring systems, even for small and poorly defined crystals.

The basic handling unit for crystallization plates is a 48-plate rack, capable of accommodating the maximum achievable daily throughput of our system. A plate crane delivers the plate to the VersaScan image processing unit, and stacks observed plates into a second rack, which is manually returned to a temperature controlled incubator. Automation of this particular plate handling step is quite urgent and easily accomplished, and given a single plate crane unit in our robotics menagerie, we assigned priority to full automation of image analysis.

2.6. Crystal harvesting and robotic diffraction screening

Crystal harvesting in suitable cryo-loops with magnetic bases has become an inexpensive and reliable de-facto standard in cryo-crystallography (Rogers, 2001). Sweeps in cryo-buffers not only provide cryo-protection, but at the same time allow introduction of heavy metals or anions such as bromide and iodide. 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 should become an increasingly interesting alternative to synchrotron based multi-wavelength methods.

Full automation of harvesting micro-manipulations appears cost prohibitive at present for all but the most affluent industrial or large facility installations, and we are currently not attempting automation of crystal harvesting in cryo-loops (although optimised crystallization plate design reduces the efforts spent in the process). If crystals become so plentiful that mounting develops into the rate limiting step, the proven success at that point may well justify further substantial investment in (or funding of) high throughput robotic crystal harvesting.

On the other hand, automated mounting of the cryo-pins on the diffractometer does greatly enhance utilization of valuable synchrotron (and lab source) beam time, and the first commercial systems are becoming available. Under the assumption that any crystal deserves screening, fast and reliable storage and mounting procedures are needed to realize high-throughput data collection for macromolecular crystallography. At the TB consortium crystallization facility, we use a sample transport and storage system developed at the Advanced Light Source (ALS) Macromolecular Crystallization Facility together with the Engineering Division of Lawrence Berkeley National Laboratory (Snell *et al.*, 2002). 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 tested and installed on ALS protein crystallography beam line 5.0.3 (Figure 3). Seven puck cassettes fit into a standard dry shipping Dewar.

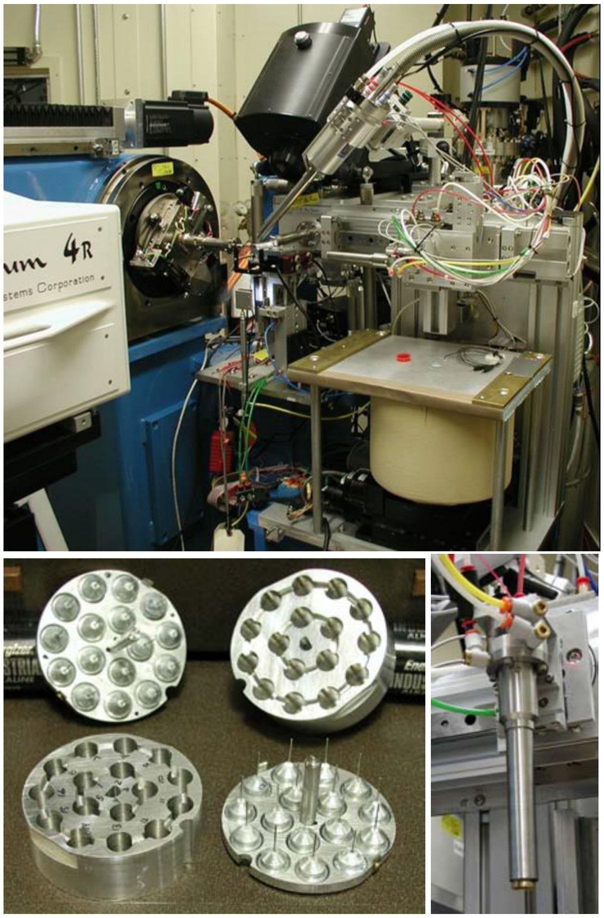

Figure 3

ALS developed automated sample mounting system. Top: Overall view of sample mounting robot in hutch of beam line 5.0.3. Bottom left: detail view of pucks, 4 of each contained in the Dewar visible at the bottom of top panel. Bottom right: 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. Sample temperatures remain at or below 110 K during the mounting and unmounting process. The loops are automatically centered on a motorized goniometer head (left side of instrument in top panel).

The mounting robot can select any of 64 samples, stored in four 16-pin pucks placed in a liquid nitrogen vessel. Mounting of a crystal with a cooled, robotic gripper takes approximately 10 seconds, during which the crystal temperature is maintained below 110 K. Centering of a crystal can be done by a user through the remote controlled goniometer head or automatically by a centering algorithm. Following initial analysis of diffraction snap-shots, the best crystal of a given protein is selected and data sets are collected. Great care is taken that in case of doubt about the space group, lower Laue symmetries are selected, and that good low resolution data are obtained, if necessary by a second, faster low resolution sweep to avoid pixel saturation. The need for good low resolution data for any phasing method (including MR) has been pointed out repeatedly, for example consult Dauter and Wilson, 2001. For the ease of model building and successful use of automated procedures, except in special cases, we do not collect data sets with resolution worse than 2.5 Å, but rather pursue additional crystallization optimisation. We estimate that our overall throughput is greater using the high resolution strategy in view of the increased difficulty to accurately build and refine low resolution models.

2.7. Automated molecular replacement

Once native data are obtained either in-house on larger crystals or at the synchrotron, availability of a homology model opens the possibility for MR phasing. The hope is that every successful MR solution might save an additional phasing experiment. If the process is sufficiently automated, the approach is justified; with the caveat 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 an acceptable free R value. Given the anticipated rise in coverage of structural folds available in the public database due to structural genomics efforts, and given innovations in the method increasing the radius of convergence for powerful MR programs (Adams *et al*., 1999, Read 2001), MR will very likely see constantly increasing use.

The fact that a backbone model delivers a weak MR solution does not yet mean a good structure will result quickly. Equally important is to subject the model, if necessary in repeated cycles, to effective bias removal techniques, as the effects of model bias can be insidious and are not easily recognized by commonly used global structure quality descriptors such as R and freeR (Kleywegt and Jones, 1997). We use a relatively simple to implement automated protocol based on MR and homology structure prediction to evaluate the potential for obtaining a reliable structure model rapidly. We identify a set of possible template structures with multiple sequence alignment tools, beginning with primary pair-wise search and subsequent multiple alignment with PSI-Blast and CLUSTALW, and retrieve them 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 (Zemla, 2002). Parallel molecular replacement searches for each of the highest scoring models using the six dimensional evolutionary search program EPMR (Kissinger, 1999) are branched to a computer cluster and the models are evaluated according to their correlation coefficient to observed data. A recent review suggests that fold recognition models, although steadily increasing in quality (Jones, 2001), still may not produce successful MR probes. While in conventional homology modelling experimental verification often is not available (or desired), the immediate feedback possible through evaluation of the model against experimental data allows 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. Side chains of the target sequence are built using SCWRL (Bower *et al*., 1997) for the best MR solution. Particularly marginal MR solutions are refined by simulated annealing torsion angle molecular dynamics using CNS (Brünger *et al*., 1998) to bring them within the convergence radius of the subsequent highly effective bias removal and map reconstruction protocol, Shake&wARP (S&W), which is our derivative of the original wARP procedure (Perrakis *et al*., 1997) which has been used successfully in automated model building (Perrakis *et al*., 2001). Subsequent to initial map calculations a single round of ARP is used to build the first set of model 'water' atoms into the S&W maps. The resulting coordinates are used in a final round of S&W to generate the electron density map to be used in the initial rebuilding. 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 data of the TB Consortium in fact have been automatically processed from a modest MR solution with a correlation coefficient of 0.32 to a high quality, bias minimized electron density map (gene rv3465, Figure 4). Automated model building efforts are rapidly progressing in a number of laboratories, and we expect to implement automated molecular replacement service for TB consortium members on a web server cluster.

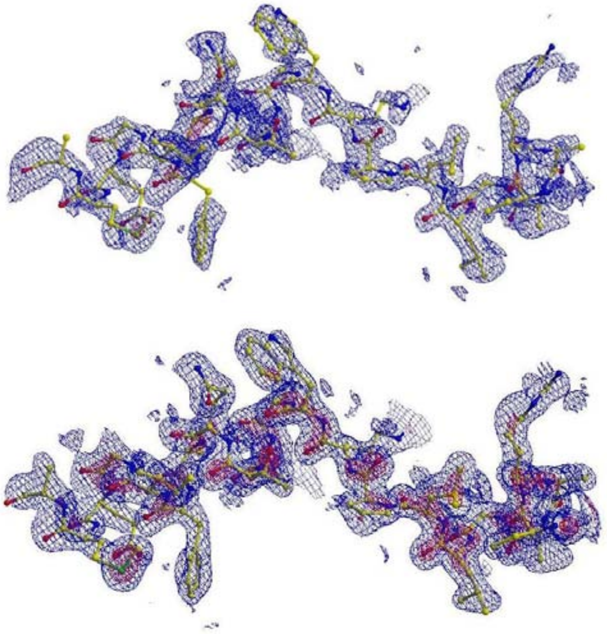

Figure 4

Electron density of Rv3465 maps, created by automated scripts directly from Scalepack output and MR homology model; blue contours at one sigma density level. Top: Refmac5 2mFo-DFc maximum likelihood maps in unmodelled C-terminal region of molecule. Bottom: same region of map, bias reduced Shake&wARP map. Note the increased clarity and connectivity of the S&W map, increasing the ease of manual (or convergence of automated) model building.

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 procedures and instrumentation, demonstrating that high throughput structure determination is possible even for small workgroups - and at a reasonable cost.

Many individuals have made contributions to the methods and material presented in this paper. We wish to acknowledge G. Meigs, C. Cork, R. Nordmeyer, E. Cornell, D. Yegian, J. Jaklevic, and J. Jin, Lawrence Berkeley National Laboratory, for their collaboration with TE and GS in the development of the crystal mounting robotics. We thank Tom Terwilliger, LANL, for his insightful management of 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) and M. Park and G. Waldo (LANL) have supplied a steady flow of proteins. Thanks are due to the members of Jim Sacchettini's group at Texas A&M University for providing numerous drug target complex structures as test cases for the automated bias removal procedures. LLNL is operated by University of California for the US DOE under contract W-7405-ENG-48. This work was funded fully by NIH P50 GM62410 (TB Structural Genomics) center grant.

References

Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1999). *Acta Cryst*. D**55**, 181-190.

Bower, M., Cohen, F. E. & Dunbrack, R. L. Jr (1997). *J. Mol. Biol.* **267**, 1268-1282.

Brünger, A. T. (1998). *Acta Cryst.* D**54**, 905-921.

Dauter, Z. & Wilson, K. S. (2001). *Principles of monochromatic data collection. International Tables For Crystallography*, Vol. F, pp. 177-195. IUCr/Kluwer Academic Publishing, Dodrecht, NL.

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. (2002). *The TB Structural Genomics Consortium: Providing a Structural Foundation for Drug Discovery. Current Drug Targets - Infectious Disorders*. In the press.

Jones, D. T. (2001). *Acta Cryst*. D**57**, 1428-1434.

Kissinger, C. R., Gelhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst*. D**55**, 484-491.

Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymology* **277**, 208-230.

Krupka, H. I., Rupp, B. & Segelke, B. W. (2002a). *Automated CRYSTOOL: A Random Screening Method for Efficient High-Throughput Crystallization.* In preparation.

Krupka, H. I., Azarani, A., Segelke, B. W, Lekin, T.P., Wright, D., Todd, P. & Rupp, B. (2002b). *Acta Cryst*. D**58**, 1523-1526.

Krupka, H. I., Segelke, B. W., Robbins, A. & Rupp, B. (2002c). *Assessment of 96 well plates for automated crystallization screening setup.* In preparation.

Luft, J. R., Wolfley, J., Jurisica, I., Glasgow, J., Fortier, S. & DeTitta, G. (2001). *J. Cryst. Growth,* **232**, 591-595.

Norvell, J. C. & Zapp-Machalek, A. (2000). *Nature Struct. Biol.* **7**, *Structural Genomics Supplement*, 931.

Perrakis, A., Sixma, T. K., Wilson, K. S. & Lamzin, V. S. (1997). *Acta Cryst*. D**53**, 448–455.

Perrakis, A., Harkiolaki, M., Wilson, K. S. & Lamzin, V. S. (2001). *Acta Cryst*. D**57**, 1445–1450.

Read, R. J. (2001). *Acta Cryst*. D**57**, 1373-1382.

Rogers, D. W. (2001). *Cryocrystallography techniques and devices. International Tables For Crystallography*, Vol. F, pp. 202-208. IUCr/Kluwer Academic Publishing, Dodrecht, NL.

Rupp, B. (2002). *Acc. Chem. Res*., *Structural Genomics Special Issue*. Submitted.

Segelke, B. W. (2001). *J. Cryst. Growth,* **232**, 553-562.

Segelke, B. W. & Rupp, B. (1998). *ACA Meeting Series*, **25**, 78.

Snell, G., Meigs, G., Cork, C., Nordmeyer, R., Cornell, E., Yegian, D., Jaklevic, J., Jin, J. & Earnest, T. (2002). *J. Synchrotron Rad.* In the press.

Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. (1999). *Nature Biotechnol.* **17**, 691-695.

Zemla, A. (2002). *Automated 3D Protein Structure Predictions Based on Sensitive Identification of Sequence Homology*. In preparation,. http://predictioncenter.llnl.gov.

Received 26 March 2002 • Accepted 7 August 2002