



Laboratory scale structural genomics

Brent W. Segelke¹, Johana Schafer¹, Matthew A. Coleman², Tim P. Lakin¹, Dominique Toppani¹, Krzysztof J. Skowronek¹, Katherine A. Kantardjieff³ & Bernhard Rupp^{1,*}

¹Macromolecular Crystallography and Structural Genomics Group, Biology and Biotechnology Research Program, P.O. Box 808, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA; ²Health Effects Genetics Division, Biology and Biotechnology Research Program, P.O. Box 808, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA; ³W.M. Keck Foundation Center for Molecular Structure, Department of Chemistry and Biochemistry, California State University Fullerton, 800 N. State College Blvd., Fullerton, CA 92834-6866, USA; *Author for correspondence (tel: +1-925-423-3273; fax: +1-925-424-3130; e-mail: br@llnl.gov)

Received 13 June 2003; accepted in revised form 22 October 2003

Key words: cloning, crystallization, expression, high throughput, TB structural genomics

Abstract

At Lawrence Livermore National Laboratory, the development of the TB structural genomics consortium crystallization facility has paralleled several local proteomics research efforts that have grown out of gene expression microarray and comparative genomics studies. Collective experience gathered from TB consortium labs and other centers involved in the NIH-NIGMS protein structure initiative allows us to explore the possibilities and challenges of pursuing structural genomics on an academic laboratory scale. We discuss our procedures and protocols for genomic targeting approaches, primer design, cloning, small scale expression screening, scale-up and purification, through to automated crystallization screening and data collection. The procedures are carried out by a small group using a combination of traditional approaches, innovative molecular biochemistry approaches, software automation, and a modest investment in robotic equipment.

Introduction

The success of the genome sequencing projects demonstrates the feasibility and utility of large scale, discovery driven biological research, and has enabled many new post-genomic research efforts at academic laboratory scale (Goulding and Perry, 2003), as well as high throughput industrial research (Blundell *et al.*, 2001; Goodwill *et al.*, 2001; Harris, 2001). The NIGMS protein structure initiative (PSI) pilot projects (Terwilliger, 2000), while intended primarily to increase the coverage of the protein fold space (Kim, 1988; Norvell and Zapp-Machalek, 2000), will also enable as yet unforeseen research. An intended derivative benefit of the PSI is to create infrastructure and drive innovation for increased capacity and decreased cost of structure determination. It can be anticipated that the resulting technological innovation

will resonate throughout the structural biology community and beyond.

To date, structural genomics pilot projects, PSI centers, international structural genomics efforts, and commercial endeavors have generated significant advancements, although the collective experience also reveals significant challenges ahead. Innovations derived from structural genomics research over the last half decade have ranged from rather simple changes that have significant impact, such as parallelization or miniaturization, to engineering solutions for automation and process development, technological achievements, and some fundamentally different ways of conducting research. A few examples of emergent technology that have impacted structural genomics are newly commercialized approaches to cloning; new expression vectors for enhanced solubility (Kapust and Waugh, 1999) and cleavage specificity (Fox

Table 1. Throughput estimates for a typical small scale structural genomics project, as implemented at the TB structural genomics crystallization facility.

Process step	Throughput estimate (target/person/day)	Currently used instruments
PCR amplification	100's	96-well pcr engine, 96-well vacuum manifold pcr cleanup
Digestion/ligation	100's	96-well pcr engine
Transformation/colony picking	100's	Mini-prep 96-well vacuum manifold miniprep ^a
<i>In vitro</i> expression screening	100's	None
<i>In vivo</i> expression screening	96	Multi-gel box
Affinity binding assay	100's	RoboPop (Novagen) ^b
Solubility screening	10's	Multi-gel box
Scale up expression	1–5	Floor shaker
Large scale purification	1–2	FPLC ^c
Crystallization screening	10's	Hydra-Plus-One (Apogent)

^a Investment in a colony picker greatly reduces tedium and errors and would increase throughput.

^b Not required, could be done by batch binding in combination with plate centrifugation.

^c Could be done with gravity affinity chromatography and other instruments.

et al., 2003; Hammarstrom *et al.*, 2002); research and development of *in vitro* transcription-translation methods (Yokoyama, 2003); emerging protocols that greatly simplify *in vivo* expression and increase yields (Studier, unpublished results); microfluidic free interface crystallization (Hansen *et al.*, 2002); and a variety of robotic instruments to automate nearly all phases of the structural genomics pipeline. New data management and laboratory information management systems (LIMS) are also being developed (Haebel *et al.*, 2001; Harris and Jones, 2002), enabling process automation and comprehensive data mining (Luft *et al.*, 2003; Rupp, 2003b), which will allow investigators to make much more rigorous and statistically sound comparisons of methods than is possible today. Investigators will have a quantified statistical basis for choosing one system or method over another, whereas these choices today are generally made on an empirical basis or simply by preference.

Results and discussion

Based on the collective experience gathered from TB consortium labs and other centers involved in the NIGMS protein structure initiative (Terwilliger, 2000), and at our facility during the past 3 years, we have undertaken process engineering efforts towards structural genomics at the laboratory scale. The primary aim of the TB structural genomics consortium crystallization (TBSGX) facility was to develop affordable, modular technology for high throughput

crystallization (Rupp *et al.*, 2002; Krupka *et al.*, 2002), and to increase our overall efficiency via process analysis (Rupp, 2003a). Here we discuss additional procedures, equipment, and infrastructure needed to carry out structural genomics projects on a laboratory scale. Despite the small size of the TBSGX group, with only four full time employee equivalents, we are assembling a complete structural genomics pipeline building from a standard molecular biology laboratory setup with only modest investment into robotics. We hope that our experiences shared in this paper could help others in developing an efficient structural genomics effort on a modest budget.

Targeting

Our philosophy of targeting in a small structural genomics group is to achieve a balance between 'conventional' hypothesis driven targets and a discovery driven component. In our 2-tiered strategy we classify the targets at each step into a promising subset and into a less likely category for success, and we pursue only the promising targets in the first round with the standard set of 'high throughput' protocols. We therefore have to accept a certain level of attrition at each step, and a sufficiently large set of carefully selected target genes is necessary. All of our gene targets have been selected from microbial pathogens and most can be classified as putative virulence factors or putative therapeutic drug targets.

For projects which are locally pursued from the beginning, targeting decisions are made either

through collaboration with local groups carrying out expression microarray experiments or comparative genomics (*Y. pestis* targets). The generic TB consortium targets for *M. tuberculosis* (MTB) have been compiled through literature search and personal interests (Schroeder *et al.*, 2002; Huang *et al.*, 2002; Goulding and Perry, 2003; Goulding *et al.*, 2002), and a large fraction of genes targeted by the consortium are aimed at populating the protein 'fold' space, as prescribed by the NIH-NIGMS. Our own MTB target list selection is based on a bioinformatics approach, which has been extensively used to support structural genomics by selecting targets for high throughput structure determination (Bertone *et al.*, 2001; Goh *et al.*, 2003; Knowles and Gromo, 2003) to obtain optimally useful solved structures. A number of clustering approaches may be used to select and prioritize targets for X-ray or NMR investigations (<http://www.structuralgenomics.org/>) and for archiving structural knowledge on experimental and predicted models of proteins (<http://presage.berkeley.edu/>). We aim to solve primarily structures of putative novel therapeutic targets, some of which (an estimated 10%) will also contribute to populating the overall fold space of the TB genome.

An immense body of information about the protein content of the proteome of *M. tuberculosis* became available with the completion of the genome sequence of the H37Rv strain in 1998 (Cole *et al.*, 1998). Genomic analysis confirmed the importance of lipid metabolism in the life of the tubercle bacillus, as well as the existence of novel protein families, PE and PPE (proline-glutamate repeats), unique to mycobacteria (Tekaiia *et al.*, 1999). To remain accurate and relevant, the annotation of the genome sequence of *Mycobacterium tuberculosis* is regularly updated using various informatics approaches (Camus *et al.*, 2002) to generate new coding sequences and classify these sequences into one of 11 functional classes (Cole *et al.*, 1998). Furthermore, comparisons may be made with other sequenced genomes, including related mycobacteria such as *M. leprae*, and the clinical strain CDC1551 to understand better how polymorphisms may be implicated in virulence (Fleishman *et al.*, 2002; Alland *et al.*, 2003).

We have used the newly annotated genome sequence (Camus *et al.*, 2002) to reexamine H37Rv coding sequences for similarities to other recently deposited or previously overlooked sequences. Standard protocols are used to perform intraproteome comparisons (Altschul *et al.*, 1990, 1997) and calculate mo-

lecular properties (<http://us.expasy.org>) to determine whether encoded proteins have a likely probability of success. Functional insights are obtained using the PROSITE (Falquet *et al.*, 2002) and InterPro (Mulder *et al.*, 2003) databases. GenTHREADER, a sequence profile-based fold recognition method (Jones, 1999; McGuffin and Jones, 2003), is also used to detect similar folds for genomic sequences or confirm absence of structural homology to known sequences.

Cloning

Cloning is carried out by conventional directed cloning methods, which yield acceptable success rates for bacterial targets. A small collection of expression vectors with compatible multicloning sites has been compiled so that with a one primer pair, a variety of expression constructs can be generated for a targeted gene. The expression vector systems in use are all T7 based, and contain compatible or engineered multicloning sites. The vector systems include the pET28 (Novagen) derived C-terminal GFP fusion (Waldo *et al.*, 1999), a modified pETBlue (Novagen) C-terminal His₆ tagged system, and a pIVEX-MBP modified cleavable N-terminal, dual tag His₆-MBP-Xa-TEV-GOI construct (Roche Biosciences). All targets are ligated into a 5' NdeI and a 3' BamHI site. A simple Perl script (derived from examples in Tisdall, 2001) parses multi-record Fasta or GenBank format sequence files and generates PCR primer pair sequences. After parsing, the software examines each gene sequence in the input for internal NdeI and BamHI sites. The software tries to substitute AseI for NdeI if an NdeI site exists internal to the gene sequence, and BglII for BamHI if a BamHI site exists internal to the gene sequence. If both NdeI and AseI or BamHI and BglII internal sites are present, the target gene is not pursued. Based on analysis of predicted or known orfs in the *Yersinia pestis* genome, less than 7% of identified targets are lost due to such internal restriction sequences. The 5' primer sequence generated by the software is simply a copy of the 5' end of the coding sequence with the start codon replaced with ATG with the 5' end further padded with nucleotides CTCGAATTCCAT (or CTCGAATTC-ATTA for AseI) to ensure efficient cleavage. The 3' primer sequence is simply a copy of the 5' end of the reverse complement of the encoding sequence, with the complement to the stop codon stripped and nucleotides GGAATTGGATCC (or GGAATTAGATCT for BglII) appended to the 5' end. Nucleotides are copied

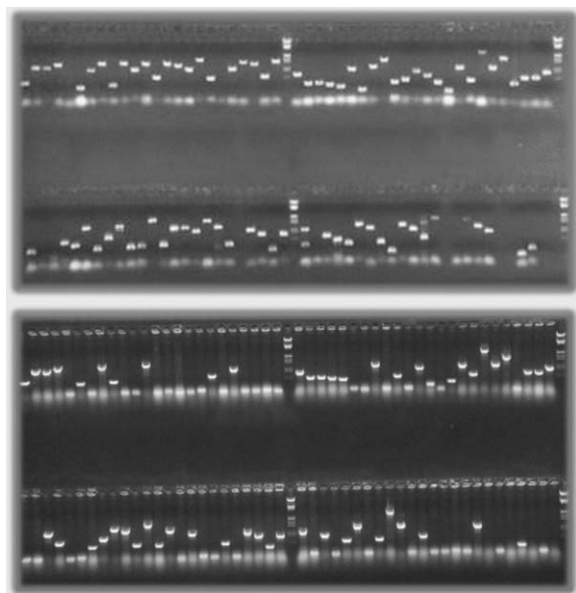


Figure 1. Ethidium bromide stained agarose gels. Top panel: 96-well PCR amplification of bacterial genes using automatically designed PCR primers. More than 90% of the targeted genes did amplify. Bottom panel: positive clones identified by colony PCR with vector specific primers flanking the multi-cloning site.

from the coding sequence (or reverse complement) to the primer sequence until T_m reaches 68 °C. For complete details, primer design scripts can be downloaded from http://porter.llnl.gov/proteomics/software_tools.

With these primers, 90.4% of targeted ORFs were successfully amplified in 96-well format with a single amplification program (Figure 1, top). Following an effort to array primers in 96-well format, PCR amplification, restriction digestion, PCR cleanup (Qiagen), ligation, and transformation to a DH5 α (DE3) amplification strain (Novagen) can all be carried out by a single individual with a thermocycler and a multipipetter in one day. Transformed cells are plated in a traditional way on LB-agar plates with antibiotics. Colonies are screened for positive clones by colony PCR with vector specific primers flanking the multi-cloning site (Figure 1, bottom), with cloning success rates approaching 90%. Positive colonies with cloned inserts of the appropriate size are rearranged and cultured in 96-well format. Clone plasmids are isolated by vacuum manifold mini-preps (Qiagen or Beckman) for archiving, sequencing, and transformation into expression strains or for use with *in vitro* translation transcription (Yokoyama, 2003). Plating and picking transformants is a significant bottleneck in

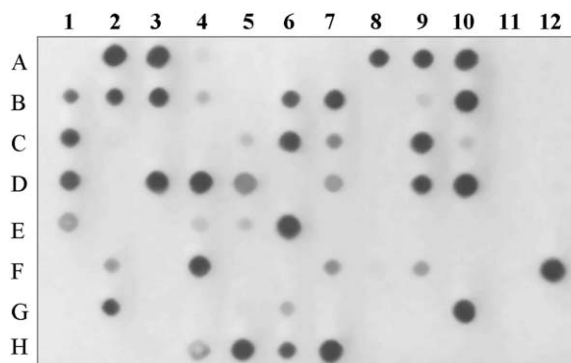


Figure 2. His-tag antibody Western plot of total *in vitro* expressed set of 96 C-terminally tagged *Yersinia pestis* clones. Full length expression is detected for 50–60% of the clones.

cloning, but an individual researcher can process multiple picks (typically 3 picks per plate) from each of 96 plates in under a day.

Expression screening

Clones are partitioned by an initial small scale expression screen into a 2-tier scheme containing a promising or not promising category similar to the one used in crystallization (Page *et al.*, 2003). Both cell free *in vitro* translation (Hino *et al.*, 2002) using the Roche Rapid Translation System (RTS) 100 *E. coli* High Yield (HY) system and conventional *in vivo* (*E. coli*) expression were evaluated.

RTS 100 *E. coli* HY expression

25 μ l *in vitro* translation (IVT) reactions are carried out in 96-well format using the RTS 100 *E. coli* HY Kit either with clone plasmid or with purified linear template PCR amplified from a clone plasmid. Expressed proteins are detected either by Western blot, using an anti-His tag antibody, or by direct fluorescent labeling of the protein with tRNA-lysine-BO-DIPY conjugate FluoroTect GreenLys (Promega) added directly to the IVT reaction (Beernink *et al.*, 2003). Ninety-six clones are expression screened by incubating at 30 °C for 3 h. IVT products are detected by dot blot. Following incubation, the contents of the IVT reactions are applied to the PVDF membrane by vacuum using a Bio-Dot blotting apparatus (Bio-Rad) and washed. The membrane is then imaged in a Packard Fluorimager (Figure 2). Typically 50–60% of clones show detectable expression and these clones

are binned into the 'promising' category. The remaining clones are binned in the unfavorable category and the corresponding gene targets are recycled for cloning and expression screening with a different construct.

The principal advantage of IVT expression screening is the ease of parallelization. With no need to monitor cell density or to induce expression, all the expression reactions can be treated the same and therefore arrayed rapidly in 96-well format. Although the correlation between expression yield determined in small-scale IVT reaction and scaled-up *E. coli* expression is quite strong for clones that express well, the general correlation of expression yields is $\sim 70\%$ (Beernink *et al.*, 2003). A unique advantage of the IVT systems is the capability to express cytotoxic proteins and to tightly control and supplement the transcription/translation reaction (Yokoyama, 2003). In practice, the convenience and advantages of IVT need to be weighed against the substantial costs.

In vivo screening

It is now feasible to carry out *in vivo* parallel expression screening as well, thanks to the advent of auto-induction methods (F. William Studier, Brookhaven National Laboratory, personal communication, studier@bnl.gov). Autoinduction methods have been developed to control the spontaneous induction observed with some batches of commercial complex media (Grossman *et al.*, 1998). Protocols have been developed for use with DE3 lysogen cell lines in combination with expression vectors for which the T7 *lac* promoter controls expression of the target gene (Studier *et al.*, 1990). For *in vivo* protein expression screening, clones are transformed in *E. coli* BL21(DE3) (Novagen) and arrayed in 96-well deep well plate (2 mL volume per well) with 250 μ L of Studier ZYP-5052 defined media. The Studier media allow growth to high densities in a standard rotary shaker-incubator. Cells are grown to saturation and auto-induced (F. William Studier, personal communication). Expression yield is assessed by SDS-PAGE electrophoresis for up to 96 samples using the Mini-PROTEAN 3 DodecaCell (BioRad). Approximately 80% of clones yield detectable full length expression (Figure 3). Freezer stocks of transformed cells can be used for both expression screening and scaled expression.

The principal advantage of *in vivo* expression screening is the cost savings compared to IVT and the

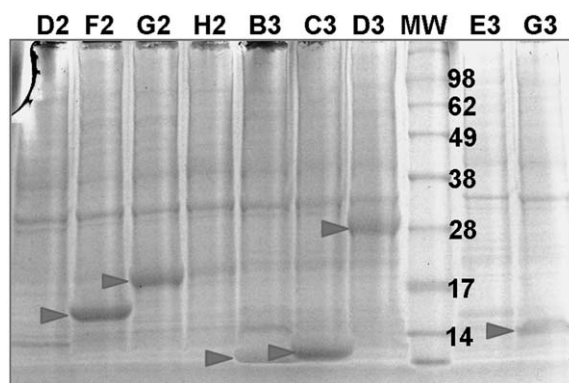


Figure 3. Auto-induced protein expression screening. Transformed clones are expressed by auto-induction (Studier, unpublished) in 250 μ L cultures and assayed for expression yield and protein size by SDS-PAGE. Approximately 80% of clones have detectable expression.

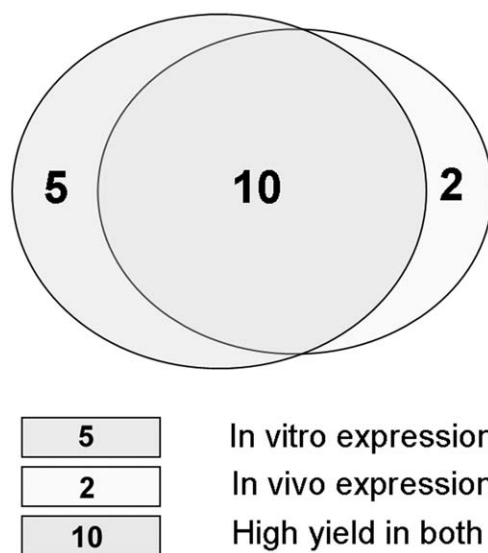


Figure 4. *In vitro* versus *in vivo* expression screening. A subset of 18 clones was screened for expression both *in vivo* and *in vitro* for comparison. For clones with high expression yield, expression was detected in both systems. Combining both approaches, all but one clone gave detectable expression.

simple scale-up with a high correlation between screening and large scale expression in *E. coli*. In a control experiment we observed a significant overlap between targets expressing in IVT screening and *E. coli* expression (Figure 4), suggesting that large scale expression in *E. coli* is in most cases a much less expensive alternative to large scale expression *in vitro*.

Protein solubility screening

Cell stocks showing detectable expression are rearrayed and grown in 2.5 mL cell culture with ZYP-5052 defined media, and expression is autoinduced. The resulting batch of cells is partitioned in 5 mL aliquots for screening against 5 different lysis buffers. The buffers are 0.5 M NaCl, 20 mM Tris-HCl pH 7.9, 20 mM imidazole, 50% glycerol (buffer1); buffer1 with 2% triton (buffer2); buffer1 with 2% NP40 (buffer3); or buffer1 with 2% Tween20 (buffer4); buffer 5 is BugBuster (Novagen). Each sample is pelleted, resuspended in lysis buffer, and sonicated. After lysis the soluble fraction is assessed for protein yield by SDS-PAGE. Greater than 90% of clones that yield detectable expression also have some detectable protein in the soluble fraction for at least one of the lysis buffers. In a comparative study of 15 targets, 13 have detectable soluble expression after lysis with buffer 3, and only 5 have detectable expression after lysis with BugBuster. All but one target gave detectable expression when all five lysis buffers were tried. In practice, buffer3 is tried first and the parallel solubility screen is reserved for cases when lysis in buffer3 does not give detectable soluble expression. Sonication is presently carried out with a Misonix ultrasonicator XL equipped with a single microtip horn, rendering lysis and solubility screening inherently low throughput, though 96 clones can be screened for soluble expression and lysis conditions in a few days by 1 or 2 persons. Sonication can be easily accelerated with the purchase of a 96-well sonication horn (Misonix). Clones that show detectable yield of soluble expression are then prioritized for scaled-up expression.

Scale-up expression and purification

Scaled expression and purification are inherently time-consuming steps compared to all preceding procedures. It is difficult to miniaturize or parallelize these steps to any significant degree, and automation brings limited throughput gains only at a high cost. Despite these limitations, if a large number of clones that yield soluble expressed proteins are pipelined, a modest sized effort can still submit a significant number of purified proteins to crystallization trials. For scaled expression, cell stocks of transformed clones yielding high amounts of soluble protein are cultured in 500 mL of cell cultures ZYP-5052 defined media, and expression is autoinduced. Cells are treated in

a standard fashion for protein preparation, except that they are resuspended in the preferred lysis buffer as determined in the initial solubility screening. Cells are lysed with an EmulsiFlex-C5 homogenizer (Avestin).

All proteins of interest are purified from the soluble fraction through the same general scheme of affinity purification followed by gel filtration. Affinity purification can be carried out using parallel small scale batch purification to assess binding and specificity of binding resin. Affinity tagged proteins that interact well with the affinity resin are purified from scaled expression by batch binding to the affinity resin, followed by gravity flow elution or elution on a low pressure chromatography instrument (BioRad). Affinity purification is followed by FPLC (Pharmacia) gel filtration. If the protein is not purified to homogeneity, gel filtration may be followed by ion exchange chromatography as well. Though gel filtration would normally be a polishing step in purification, ion exchange at the last stages has the advantage of re-concentrating protein while providing another step of separation. If proteins of interest yield to the standard protocols, several proteins a week can be produced using a single FPLC instrument.

The final purified protein is exchanged into minimal buffer for crystallization either by dialysis in the final buffer, buffer exchange in Centricon or Amicon concentrators, or by desalting on a desalting column (Amersham). The suitable protein concentration for crystallization is determined using a prescreening procedure whereby the protein stock is combined with a range of concentrations for each of three precipitating agent reagent classes (alcohol, salt, and PEG). If few or none of the reactions shows precipitation, the protein needs to be further concentrated. Precipitation in several but not all conditions indicates a concentration appropriate for crystallization screening.

Combinatorial crystallization screening

Crystallization screening is substantially automated and the entire process is developed around the automated design of crystallization screens using the CRYSTOOL program (Segelke and Rupp, 1998). By considering crystal screening as a sampling problem, we have previously demonstrated by probability theory the inherent efficiency of CRYSTOOL screening (Segelke, 2001). With CRYSTOOL we are able to generate any number of random combinations of

crystallization conditions from a large set of stock solutions and we have interfaced CRYSTOOL to an automated liquid-handling system (Packard, MPII-HT). The detailed design of and the philosophy behind our high throughput crystallization pipeline have been described previously (Rupp, 2003a; Rupp *et al.*, 2002) and are summarized below.

Cocktail production

CRYSTOOL provides automated design of novel and efficient crystallization screens by random combination from ~ 90 stock reagents. The program writes runtime instructions for the Packard Instruments MPII liquid handling robot, which produces crystallization screens from stock reagents in 96-well format (1.5 mL BioBlocks) with a capacity of ~ 10 blocks/day. The Packard is relatively fast, has eight independently actuated tips with variable span, and a liquid level sensing feature. The interfacing software allows extensive runtime control with full random access to all positions on the deck of the robot. The positional and volume precision, however, are not adequate for dispensing sub microliter volumes reliably, so it is impractical to use the MPII for the drop setups in 96-well plates, particularly at less than 3 μL per drop.

Crystallization setup with the Hydra-Plus-One

Sitting-drop experiments are set up in 96-well IntelliPlates[®] (Robbins Enterprises) using a Hydra-Plus-One robot co-developed with Apogent Discoveries (Krupka *et al.*, 2002). The Hydra-Plus-One can deliver sub microliter volumes reproducibly and with good positional precision. The Hydra-Plus-One has a significant speed advantage compared to the 8-tip MPII when setting up 96-well crystallization plates. The 96-syringe dispense head allows a single mother-daughter transfer of the pre-arrayed cocktails. An attached, additional single channel Innovadyne microsolenoid dispenser delivers protein to each of 96 drops in the crystallization plate with high precision, in just over 1 minute, with reasonably low drop volumes of 500 nL + 500 nL.

To further miniaturize the dispensing volumes we are co-developing with Innovadyne an affordable 96-channel contactless dispenser with additional one to eight channels for rapid and precise dispensing of 100–50 nL cocktail and protein drops. It is expected that the practically achievable minimum drop size will no longer be determined by the instrument pre-

cision, but rather by requirements for rapid plate sealing or humidity control (Santarsiero *et al.*, 2002).

Crystallization in the IntelliPlate

We co-developed the multi-purpose IntelliPlate 96-well sitting drop crystallization plate with Art Robbins Enterprises to incorporate several features unavailable with other SBS format crystallization plates at the time. The IntelliPlate was designed to hold up to 250 μL of reservoir solution and to accommodate a large range of drop volumes for screening and optimization. There are two positions for drop setup on the shelf of each reservoir in the plate: a semi-spherical depression that holds up to 2.8 μL , and a bathtub shaped depression that holds up to 8 μL . The smooth depressions provide for convenient crystal harvesting, and prevent drops from creeping into corners and edges, as frequently observed with square and flat-bottomed wells. The adverse optical properties of a round shape acting as a lens are accounted for in the IntelliPlate by a matching counter lens designed into the underside of each drop depression. Finally, the IntelliPlate is designed with a wider rim around each reservoir and a flush edge around the perimeter of the plate to facilitate sealing and easy stacking.

Automated imaging and image analysis with CRYSFIND

Based on a conservative throughput of 10 proteins screened per day at 288 experiments per protein, and a viewing schedule of 6 times through a 6-month lifetime of a plate, 17280 experiments per day on average need to be viewed, scored and recorded in the crystallization database (Rupp *et al.*, 2002). We thus developed an automatic CCD imaging station with integrated crystal recognition software. Viewing is accomplished using a basic plate handling robot fully integrated with the VersaScan imaging system, designed in collaboration with Velocity11 in Palo Alto. The system is capable of imaging a 96-well plate in less than two minutes, producing one-mega-pixel black and white well images, which are processed on a dedicated dual CPU computer. Crystal detection software, CRYSFIND, currently under commercial evaluation and available as a stand-alone, customizable package, can classify images of crystallization experiments with 95% accuracy (compared to a human observer) for presence or absence of crystals.

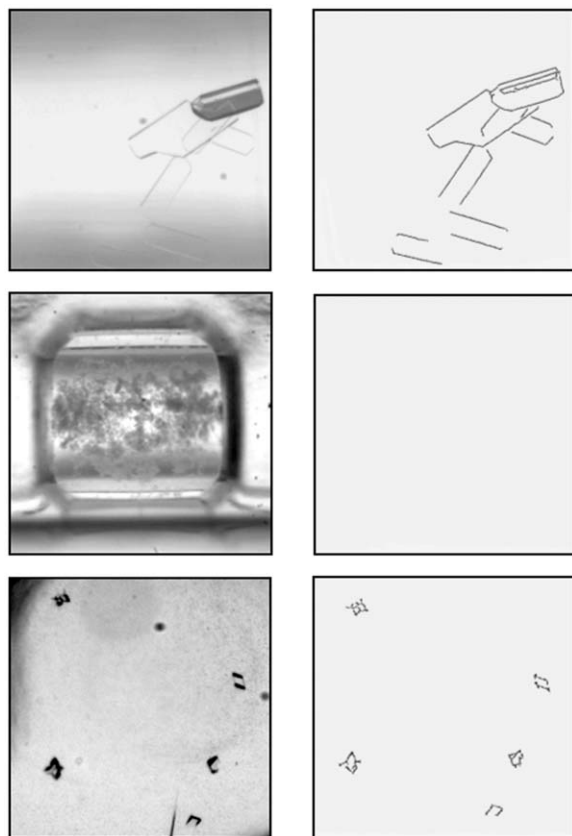


Figure 5. TB Protein Crystals. Pairs of original images (left column) and images containing extracted crystal features (right column). The CRYSFIND software is capable of detecting both very small and big crystals even in shadows and low-contrast conditions with default settings. Top row: a human user might miss a 'ghostly' crystal such as at the bottom left of the well, but the software succeeds in finding both edges, and grouping them as one big crystal. Middle row: the absence of features detected in the high-contrast, heavy precipitate demonstrates that a low percentage of false positives are expected. Bottom: even for small and irregular crystals, the extracted features are clear and allow reliable automated crystal scoring.

CRYSFIND also classifies each crystal individually in terms of size and overall quality. In first double blind trials CRYSFIND detects crystals with a 5% false positive and 5% false negative rate compared to a human observer. The software was designed to address a number of commonly encountered challenges (Jurisica *et al.*, 2001; Luft *et al.*, 2001; Wilson, 2002; Spraggon *et al.*, 2002) in crystal image recognition (Figure 5). In particular, its detection algorithm success does not depend on the picture contrast and is insensitive to shadows. A detailed description and

evaluation of the algorithms used in CRYSFIND will be given separately (Toppani *et al.*, in preparation).

The CRYSFIND crystal detection software automatically generates e-mail reports of its scores, which are then visually verified. Images and results are stored in the crystallization database, which is integrated with scripts to form a customizable basic LIMS that maintains the plate viewing schedule, notifies group members of plates to be viewed on a given day, and alerts the TB consortium members to newly discovered crystals.

The expensive path to full integration

The ultimate goal of full automation would be a walk-away system that integrates its components (for example, cocktail mixer, plate setup, sealer, imager, and plate storage silo). Such systems have been built, albeit at considerable cost (Hosfield *et al.*, 2003). We have argued that it may be sufficient and more cost effective for a smaller lab to implement plate handling only in highly redundant steps such as viewing plates or supplying labware to robots (Rupp, 2003a). We acquired a rather simple and affordable robotic arm (PlateCrane, Hudson Control Group) for these purposes. The PlateCrane has a basic software interface that can be accessed at a rather low level, providing a fair amount of versatility, but one needs extensive programming experience to realize the full potential of this device. The commercial user software has now been completely replaced with a custom interface and the PlateCrane has been integrated with the imaging system. Even after system integration, managing and handling the plates for scheduled viewing remains a labor intensive step. A Motoman SV3-J robot on a linear track has been purchased and will be integrated with a plate hotel and the imaging system within an environmental enclosure. Developing the walk-away automation for plate viewing represents a significant investment, but it relieves a major draw on personnel resources.

Automated sealing remains one of the steps hampering full automation of our plate-based crystallization setup, an issue that microbatch or microfluidic screening avoid (D'Arcy *et al.*, 2003; Hansen *et al.*, 2002) – although at the price of much more difficult harvesting. Thermal sealers are inappropriate for sealing of currently available crystallization plates, providing few options for clear thermal seals and resulting in destructive temperature rise at the protein drop (data not shown). Crystallization plates and

deep-well plates filled with crystallization screens are currently sealed with adhesive seal by hand or with a home-built pneumatic pressure sealer. A Brandel sealer suitable for full integration is currently under evaluation.

Success rates and optimization strategies

Despite automation, miniaturization, and parallelization of many steps, solubility screening, scaled-up expression, and purification remain manual and serial processes limiting total throughput of purified proteins into crystallization trials to no more than two proteins per week per investigator. To date, 242 samples representing 118 different microbial and human proteins have been processed through our crystallization pipeline.

Confirmed crystals have been reported for 37 of the 118 proteins screened, and 39 more proteins yielded marginal results where their diffraction status could not be verified. Six of the confirmed crystals yielded high resolution data sets without any need for optimization, and 15 data sets have been collected in total. Results are difficult to compare to expected values, as many proteins are supplied to the facility by individual investigators only once the first trials were unsuccessful or indicated complications, and in some cases no material is available for optimization. A reasonable projection is that 1/3 to 1/2 of protein targets that persist into crystallization trials indeed crystallize, and at least 10 to 20% of those yield diffraction quality crystals from initial screens. Diffraction screening and optimization is a significant bottleneck and currently at the forefront of the effort. The current strategy for optimization of the 2/3 of crystallizing proteins that do not immediately yield diffraction quality crystals is to generate fine screens and additive screens. This strategy often leads to better crystals, but the number of experiments required is difficult to predict. With increasing amounts of data available, predictive optimization models are being developed to maximize the likelihood of crystallization success (Jurisica *et al.*, 2001; Hennessy *et al.*, 2000; Kimber *et al.*, 2003; Page *et al.*, 2003; Rupp, 2003b). During the remaining 2 years of the NIH PSI project we expect that more than 500,000 CRYSTOOL random crystallization experiments will have been set up, and we estimate crystallization of about 400 proteins *via* random screening. The records from these combinatorial screening experiments will provide a comprehensive, densely populated, and unbi-

ased database of the crystallization parameter space. Analysis of these data will lead to custom design of more efficient screens, and provide the basis for a sophisticated approach to optimization (Rupp, 2003b).

Diffraction screening and structure determination

We have outlined the procedures and described some robotics used in our process in previous publications (Rupp *et al.*, 2002; Rupp, 2003a). As a rule, we harvest and cryo-mount every viable crystal on a standard Hampton pin, and store the crystals in a storage and mounting system developed at beam line 5.0.3. of the ALS (Snell *et al.*, 2004). The methods we generally use for high throughput protein crystallography have been reviewed (Heinemann *et al.*, 2001; Lamzin and Perrakis, 2000; Goodwill *et al.*, 2001), and we have developed our own protocols for automated molecular replacement solution (Kantardjieff *et al.*, 2002; Reddy *et al.*, 2003; Rupp, 2003a). At the throughput we (and most other Structural Genomics labs, <http://targetdb.pdb.org/>) are currently achieving, structure solution is not a rate limiting step. Most of our recent synchrotron time has been via public beam line applications, and to optimize screening throughput we only work at robotics equipped beam lines (ALS 5.0.2, 5.0.3).

Instrumentation

The design philosophy at the TBSGX facility has been to develop modular and affordable robotics. We set a price limit of US \$50–120k for each component in our modular process pipeline. We have also argued that full automation of all steps may not be necessary, and a process review should be conducted (Rupp, 2003a; Hillier and Lieberman, 2000). For example, it may be sufficient and more cost effective for a smaller lab to implement plate handling only in limited, highly repetitive operations such as plate imaging, or parallel PCR and expression screening setup.

We suggest the following minimum instrumentation to carry out structural genomics on a laboratory scale, beyond what would be found in a typical molecular biology and protein biochemistry lab: a Hydra (Apogent), or other multichannel pipetting system for parallel reagent transfers and rearranging; a Hydra-Plus-One (Apogent) or equivalent for setup of crystallization plates (Krupka *et al.*, 2002; Luft *et al.*, 2003); and an automated crystallization plate imaging system (Velocity11VersaScan or equivalent). In

addition, any versatile liquid handling system, like the Packard MPII, enables a much greater level of automation for many tedious and error prone procedures such as the mixing of custom crystallization screens or automated PCR setup. Currently only fixed pre-mixed sparse matrix (Jancarik and Kim, 1991) or grid screens (McPherson, 1982) can be purchased in 96-well format (Hampton Research). With the suggested instrumentation in place, the remaining most tedious and difficult to parallelize procedures are colony picking and protein purification other than affinity capture. There are a number of instruments available for automated colony picking, but parallel or fully automated high throughput purification requires a considerable investment.

Data management for structural genomics at a laboratory scale is a non-trivial issue. There is considerable informational entropy buildup from failures, uncaptured data, inaccessibility of data, unsuitable spreadsheet formats, and the sheer quantity of data and different and partly remote data entry points. A few LIMS systems for crystallographic structure determination are publicly available (Haebel *et al.*, 2001; Harris and Jones, 2002) and a rather extensive LIMS system is currently built for the TB Structural Genomics Consortium at UCLA (M. Parag, T. Holton, D. Pal *et al.*, personal communication).

Conclusions

The pursuit of structural genomics on a laboratory scale (three to five full time employees) appears viable, given a broad skill mixture (and cross-training) of personnel, and a modest but well-placed investment in robotics. While there is a significant cost to develop full walk-away automation solution, strategic investment in a few capital purchases can greatly enhance the capacity of a modest size structural genomics effort. The key differences in the structural genomics approach compared to traditional structural biology are parallelization, screening and decision making at intermediate steps, miniaturization, and an actuarial view that one can anticipate, plan for, and accept losses. In a consortium organization, a great diversity of approaches can be pursued, each on a modest scale, and the collective experience captured at a central database. This in turn would lead to further advancements, as methods could be more rigorously compared.

Acknowledgements

The TB consortium cloning and protein production facilities under J. Perry, C. Gouling, D. Eisenberg (UCLA), T. Terwilliger, M. Park, C.-Y. Kim and G. Waldo (LANL), have supplied a steady flow of proteins. P. Malik, T. Holton, D. Pal and co-workers have developed the TB consortium web site and database at UCLA. Christa Prange, Christine Sanders, Peter Beernink, and Brian Souza contributed to the development of *in vitro* expression screening methods. B.R. and K.A.K. thank Jim Sacchettini, Texas A&M University, for support during their sabbatical leave. 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. Work on *Y. pestis* targets was partially funded by LLNL LDRD 01-ERD-045.

References

- Alland, D., Whittam, T.S., Murray, M.B., Cave, M.D., Hazbon, M.H., Dix, K., Kokoris, M., Duesterhoeft, A., Eisen, J.A., Fraser, C.M. and Fleishman, R.D. (2003) *J. Bacteriol.*, **185**, 3392–3399.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
- Beernink, P.T., Segelke, B.W. and Coleman, M.A. (2003) *Biochemistry*, **1**, 4–5.
- Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T. and Gerstein, M. (2001) *Nucleic Acids Res.*, **29**, 2884–2898.
- Blundell, T.L., Jhoti, H. and Abell, C. (2001) *Nat. Rev. Drug Discov.*, **1**, 45–54.
- Camus, J.-C., Pryor, M.J., Medigue, C. and Cole, S.T. (2002) *Microbiology*, **148**, 2967–2973.
- Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V. et al. (1998) *Nature*, **393**, 537–544.
- D'Arcy, A., MacSweeney, A., Stihle, M. and Haber, A. (2003) *Acta Crystallogr.*, **D59**, 396–399.
- Falquet, R.D., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K. and Bairoch, A. (2002) *Nucleic Acids Res.*, **30**, 235–238.
- Fleishman, R.D., Alland, D., Eisen, J., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., Hickey, E., Kolonay, J.F., Nelson, W.C., Umayam, L.A., Ermolaeva, M., Salzberg, S.L., Delcher, A., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, A., Bishai, W., Jacobs, W.R., Jr., Venter, J.C. and Fraser, C.M. (2002) *J. Bacteriol.*, **184**, 5479–5490.
- Fox, J.D., Routzahn, K.M., Bucher, M.H. and Waugh, D.S. (2003) *FEBS Lett.*, **537**, 53–57.

- Goh, C.S., Lan, N., Echols, N., Douglas, S.M., Milburn, D., Bertone, P., Xiao, R., Ma, L.C., Zheng, D., Wunderlich, Z., Acton, T., Montelione, G.T. and Gerstein, M. (2003) *Nucleic Acids Res.*, **31**, 2833–2838.
- Goodwill, K.E., Tennant, M.G. and Stevens, R.C. (2001) *Drug Discov. Today*, **6(15)**, S113–S118.
- 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. and Park, M.S. (2002) *Curr. Drug Targets - Infectious Disorders*, **2**, 121–141.
- Goulding, C.W. and Perry, J.L. (2003) *J. Struct. Biol.*, **142**, 133–143.
- Grossman, T.H., Kawasaki, E.S., Punreddy, S.R. and Osburne, M.S. (1998) *Gene*, **209**, 95–103.
- Haebel, P.W., Arcus, V.L., Baker, E.N. and Metcalf, P. (2001) *Acta Crystallogr.*, **D57**, 1341–1343.
- Hammarstrom, M., Hellgren, N., vanDenBerg, S., Berglund, H. and Hard, T. (2002) *Protein Sci.*, **11**, 313–321.
- Hansen, C.L., Skordalakes, E., Berger, J.M. and Quake, S.R. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 16531–16536.
- Harris, M. and Jones, T.A. (2002) *Acta Crystallogr.*, **D58**, 1889–1891.
- Harris, T. (2001) *Drug Discov. Today*, **6(22)**, 1148.
- Heinemann, U., Illing, G. and Oschkinat, H. (2001) *Curr. Opin. Biotechnol.*, **12**, 348–354.
- Hennessy, D., Buchanan, B., Subramanian, D., Wilkosz, P.A. and Rosenberg, J.M. (2000) *Acta Crystallogr.*, **D56**, 817–827.
- Hillier, F.S. and Lieberman, G.J. (2000) *Introduction to Operations Research*, McGraw-Hill, New York, NY.
- Hino, M., Shinohara, Y., Kajimoto, K., Terada, H. and Baba, Y. (2002) *Protein Expr. Purif.*, **24**, 255–259.
- Hosfield, D., Palan, J., Hilgers, M., Scheibe, D., McRee, D.E. and Stevens, R.C. (2003) *J. Struct. Biol.*, **142**, 207–217.
- Huang, C.C., Smith, C.V., Glickman, M.S., Jacobs, W.R., Jr. and Sacchettini, J.C. (2002) *J. Biol. Chem.*, **277**, 11559–11569.
- Jancarik, J. and Kim, S.-H. (1991) *J. Appl. Crystallogr.*, **24**, 409–411.
- Jones, D.T. (1999) *J. Mol. Biol.*, **287**, 797–815.
- Jurisica, I., Rogers, P., Glasgow, J.I., Fortier, S., Luft, J.R., Wolfley, J.R., Bianca, M.A., Weeks, D.R. and DeTitta, G.T. (2001) *IBM Systems J.*, **40**, 248–264.
- Kantardjiev, K.A., Höchtel, P., Segelke, B.W., Tao, F.M. and Rupp, B. (2002) *Acta Crystallogr.*, **D58**, 735–743.
- Kapust, R.B. and Waugh, D.S. (1999) *Protein Sci.*, **8**, 1668–1674.
- Kim, S.-H. (1988) *Nat. Struct. Biol. Suppl.*, **5**, 643–645.
- Kimber, M.S., Vallee, F., Houston, S., Necakov, A., Skarina, T., Evdokimova, E., Beasley, S., Christendat, D., Savchenko, A., Arrowsmith, C.H., Vedadi, M., Gerstein, M. and Edwards, A.M. (2003) *Proteins*, **51**, 562–568.
- Knowles, J. and Gromo, G. (2003) *Nat. Rev. Drug Targets*, **2**, 63–69.
- Krupka, H.I., Rupp, B., Segelke, B.W., Lekin, T.P., Wright, D., Wu, H.-C., Todd, P. and Azarani, A. (2002) *Acta Crystallogr.*, **D58**, 1523–1526.
- Lamzin, V.S. and Perrakis, A. (2000) *Nat. Struct. Biol.*, **7**, 979–981.
- Luft, J.R., Collins, R.J., Fehrman, N.A., Lauricella, A.M., Veatch, C.K. and DeTitta, G.T. (2003) *J. Struct. Biol.*, **142**, 170–179.
- Luft, J.R., Wolfley, J., Jurisica, I., Glasgow, J., Fortier, S. and DeTitta, G.T. (2001) *J. Crystal Growth*, **232**, 591–595.
- McGuffin, L.J. and Jones, D.T. (2003) *Bioinformatics*, **19**, 874–881.
- McPherson, A. (1982) *Preparation and Analysis of Protein Crystals*, Krieger Publishing Company, Malabar, FL.
- Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A. et al. (2003) *Nucleic Acids Res.*, **31**, 315–318.
- Norvell, J.C. and Zapp-Machalek, A. (2000) *Nat. Struct. Biol. Suppl.*, **7**, 931.
- Page, R., Grzechnik, S.K., Canaves, J.M., Spraggon, G., Kreusch, A., Kuhn, P., Stevens, R.C. and Lesley, S.A. (2003) *Acta Crystallogr.*, **D59**, 1028–1037.
- Reddy, V., Swanson, S., Segelke, B., Kantardjiev, K.A., Sacchettini, J.C. and Rupp, B. (2003) *Acta Crystallogr.*, **D59**, 2200–2210.
- Rupp, B. (2003a) *Acc. Chem. Res.*, **36**, 173–181.
- Rupp, B. (2003b) *J. Struct. Biol.*, **142**, 162–169.
- Rupp, B., Segelke, B.W., Krupka, H.I., Lekin, T.P., Schafer, J., Zemla, A., Toppani, D., Snell, G. and Earnest, T.E. (2002) *Acta Crystallogr.*, **D58**, 1514–1518.
- Santarsiero, B.D., Yegian, D.T., Lee, C.C., Spraggon, G., Gu, J., Scheibe, D., Uber, D.C., Cornell, E.W., Nordmeyer, R.A., Kolbe, W.F., Jin, J., Jones, A.L., Jaklevic, J.M., Schultz, P.G. and Stevens, R.C. (2002) *J. Appl. Crystallogr.*, **35**, 278–281.
- Schroeder, E.K., de Souza, O.N., Santos, D.S., Blanchard, J.S. and Basso, L.A. (2002) *Curr. Pharm. Des.*, **3**, 197–225.
- Segelke, B. and Rupp, B. (1998) *ACA Meeting Series*, **25**, 78.
- Segelke, B.W. (2001) *J. Crystal Growth*, **232**, 553–562.
- Snell, G., Cork, C., Nordmeyer, R., Cornell, E., Meigs, G., Yegian, D., Jaklevic, J., Jin, J., Stevens, R.C. and Earnest, T.E. (2004) *Structure*, **12**, 1–20.
- Spraggon, G., Lesley, S.A., Kreusch, A. and Priestle, J.P. (2002) *Acta Crystallogr.*, **D58**, 1915–1923.
- Studier, F.W., Rosenberg, A.H., Dunn, J.J. and Dubendorff, J.W. (1990) *Methods Enzymol.*, **185**, 60–89.
- Tekaia, F., Gordon, S.V., Garnier, T., Brosch, R., Barrell, B.G. and Cole, S.T. (1999) *Tubercle Lung Dis.*, **79**, 329–342.
- Terwilliger, T.C. (2000) *Nat. Struct. Biol. Suppl.*, **7**, 935–939.
- Tisdall, J. (2001) *Beginning Perl for Bioinformatics*, O'Reilly and Associates, Sebastopol, CA.
- Waldo, G.S., Standish, B.M., Berendzen, J. and Terwilliger, T.C. (1999) *Nat. Biotechnol.*, **17**, 691–695.
- Wilson, J. (2002) *Acta Crystallogr.*, **D58**, 1907–1914.
- Yokoyama, S. (2003) *Curr. Opin. Chem. Biol.*, **7**, 39–43.