Structural Genomics of Minimal Organisms: Pipeline and Results

Sung-Hou Kim*, Dong-Hae Shin^, Rosalind Kim^, John-Marc Chandonia^, and Paul Adams^

^Berkeley Structural Genomics Center, Lawrence Berkeley National Laboratory, and *Department of Chemistry, University of California, Berkeley, CA, 94720

I. Introduction

Mission: The Protein Structure Initiative (PSI) of U. S. National Institutes of Health (NIH) aims to obtain structural information on all proteins derivable from their DNA sequences (http://www.nigms.nih.gov/psi/). The objective of the pilot phase (PSI-1) is summarized as follows: (1) to perform pilot studies to develop high-throughput methods and protocols to proceed from cloning to structure determination for representatives of diverse protein-sequence families with no sequence similarities to proteins of known structures, (2) to identify critical areas and steps for further development to achieve a high-throughput operation, and (3) to obtain the metrics for assessing the magnitude and scale required for the production phase of PSI (PSI-2) to achieve the overall PSI objective of a comprehensive coverage of the protein structure space.

Objective: In the pilot phase, the Berkeley Structural Genomics Center (BSGC) set the goal of obtaining structural information of a near complete set of all soluble proteins in two related minimal organisms (the pathogens**,** *Mycoplasma pneumoniae (MP)* and *M. genitalium (MG),* with ~700 and ~500 genes, respectively).

Pipeline: To achieve our mission, we have developed methods and protocols to automate or parallelize many processes from cloning the target genes to structure determination. Over all pipeline schemes for the single-path approach used in the initial two year period and the multiple-path approach used for the rest of the PSI-1 period are shown in Fig. I.1.

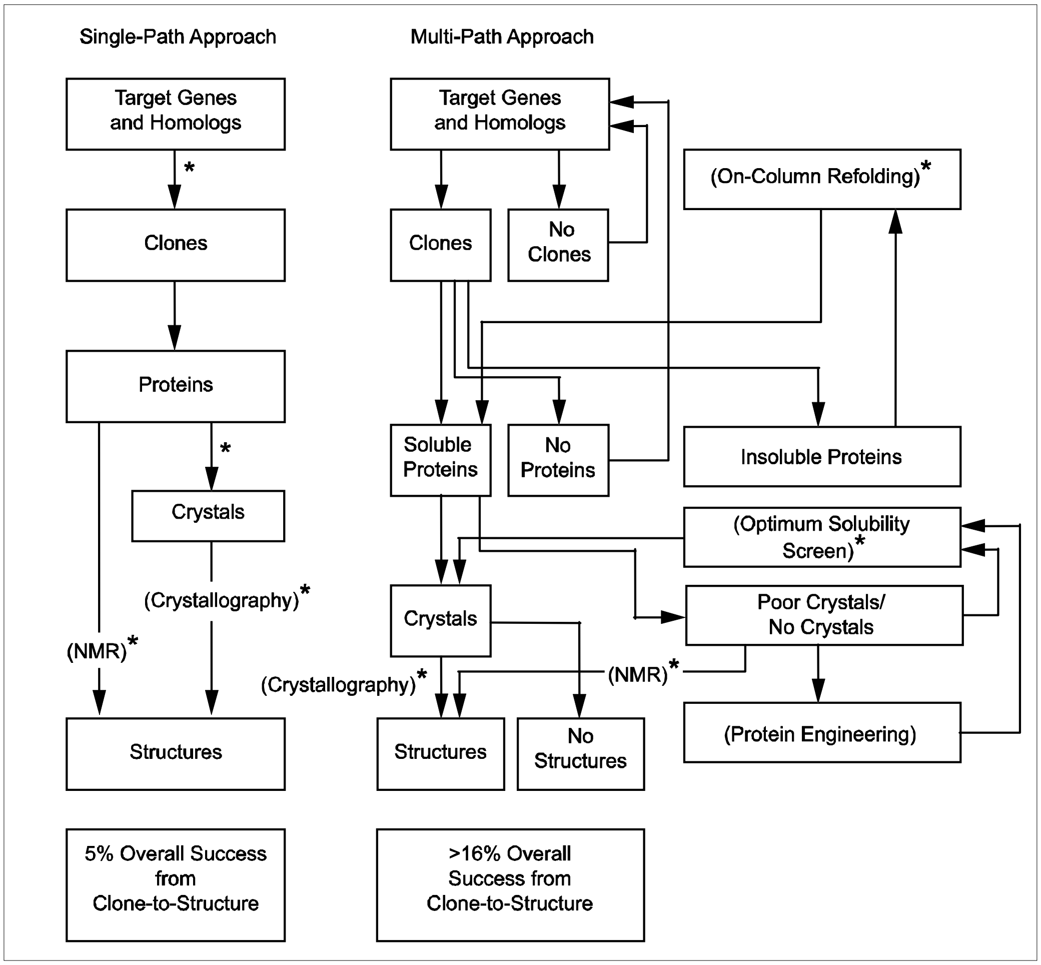

Fig. 1. Single-Path Approach vs. Multi-Path Approach for soluble proteins. A large number of target genes and their homologs were selected (see Target Selection below), and the majority of them could be cloned. Of these, some were over-expressed as proteins in soluble form, as protein aggregates, or as insoluble inclusion bodies. In the single-path mode, only the soluble proteins were screened for crystallization or NMR studies, and of these, only some yielded structures. The overall success rate for the single-path approach was about 5%. By contrast, in the multi-path mode, those clones not expressing, or under-expressing could be re-cloned with different constructs and/or into different vectors to obtain additional over-expressing clones; proteins that aggregate or could not be concentrated underwent Optimum Solubility Screening (see below) to find optimum conditions in which they were soluble and homogeneous. Those proteins that were insoluble underwent an On-column Refolding process (see below). Proteins that were chemically and conformationally homogeneous were used for NMR or crystallization studies. BSGC experience shows that for this multi-path approach the overall success rate increased to about 16%. *Processes which are automated or semi-automated.

II. Metrics and Lessons Learned

Based on BSGC's results during the PSI-1 period, we learned the metrics and lessons required for a structural genomics approach for a large scale structure determination effort, some were predicted and others were unexpected and surprising. Although some details may be different from those of other PSI-1 centers, the general conclusions of metrics and lessons are expected to be valid. They are summarized below:

1. The steps required to proceed from cloning a gene encoding a protein to determining its 3D structure can be divided into two distinct categories: (1) those where the underlying science and technologies are well understood, thus, *automatable* by instrumentation or programming, and (2) those where the underlying science is only partially known and the outcome of the processes are unpredictable. The most practical approach for steps of the second category is *multi-variable screenings***.**
2. The *single-path (low-hanging fruit) approach* (Fig. I.1), whereby, for a large number of diverse genes, one single optimized path is taken from cloning to structure determination, has less than a 5% success rate on average in discovering structures of "unique" proteins, the proteins without sequence homology to those of known structures in the Protein Data Bank (PDB) (Berman et al., 2000).
3. The *multi-path approach* (Fig. I.1), where feedback loops and multi-factor screenings are employed for one or more critical steps in the path for challenging proteins that fail by a single-path approach, has a 16% or higher success rate for discovering the structure of unique proteins.
4. Approximately half of the structures of unique proteins revealed a new fold, and the remaining half are "remote homologs" (similar structure without sequence similarity) of known folds.
5. Approximately 2/3 of the structures of "hypothetical proteins" (proteins that have no sequence homologs among the proteins of known function) infer one or a few possible molecular functions that could be experimentally tested.
6. The protein fold space can be mapped in three-dimensional space based on pair-wise structural similarities (Hou et al., 2003, 2004), thus, providing a platform for representing all protein structures, "the protein structure universe".

III. Selection of Target Proteins for High Throughput Structural Characterization

Method: A structural genomics target is a protein that is selected to determine its 3-D structure. BSGC targets during the PSI-1 period include *Mycoplasma* proteins as well as their sequence homologs from other prokaryotes. In general, all rounds of target selection involved three common steps. Since almost all *MG* genes have their homologs in *MP*, we started each step with the set of 677 *MP* ORFs described in the original annotation of the genome (Himmelreich et al., 1996). Each ORF was then augmented with a family of homologs from available, fully sequenced prokaryotic genomes to make a target set. First, all target sets recognizably homologous to proteins of known structure were removed from further consideration. Next, target sets of proteins that were predicted to be unsuitable for high-throughput study (e.g., those with predicted transmembrane helices) were eliminated. Finally, specific targets were chosen from among proteins in the remaining target sets. The number of targets chosen per family, or *parallelism*, varied amongst selection rounds. A flow diagram of target selection for a sample round is shown in Fig. 2.

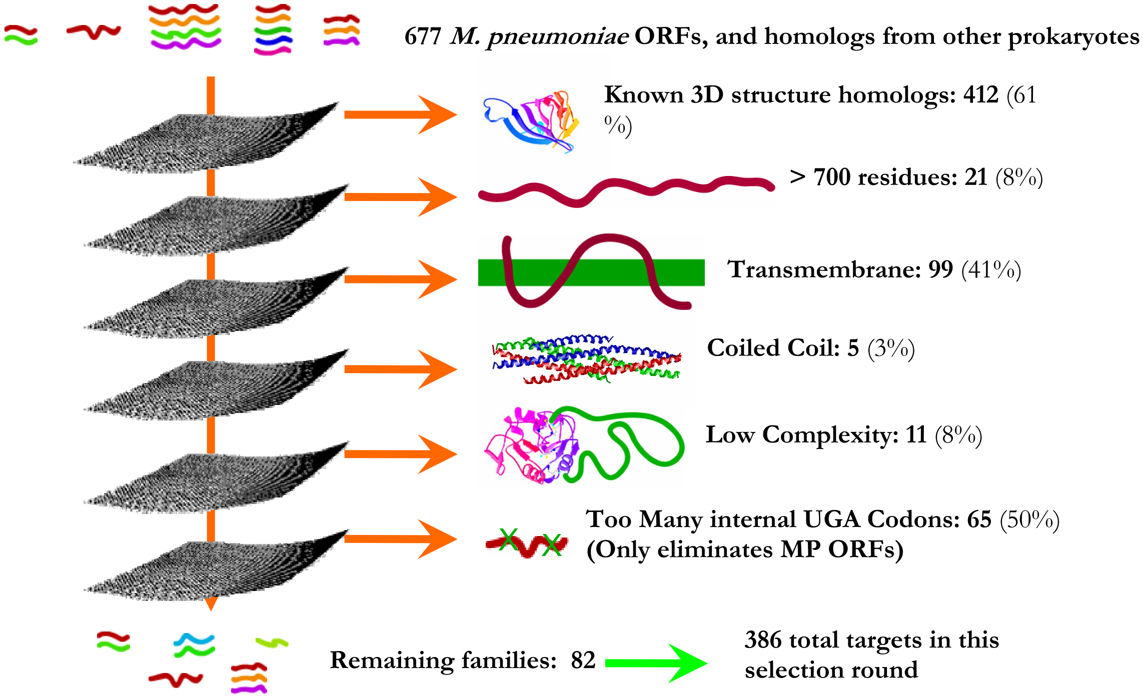
677 *M. pneumoniae* ORFs, and homologs from other prokaryotes
Known 3D structure homologs: 412 (61 %)
> 700 residues: 21 (8%)
Transmembrane: 99 (41%)
Coiled Coil: 5 (3%)
Low Complexity: 11 (8%)
Too Many internal UGA Codons: 65 (50%)
(Only eliminates MP ORFs)
Remaining families: 82
386 total targets in this selection round

Fig. 2. Target selection scheme for *Mycoplasma* genes used at BSGC. Criteria of "filtering" were changed among different rounds of target selection.

Databases: The following databases were used in selection of targets:

1. MP: we started each step with the set of 677 *MP* ORFs described in the original annotation of the genome (Himmelreich et al., 1996).
2. knownstr: a database of sequences from proteins of known structure. This database contained sequences of proteins released by PDB, sequences of proteins deposited in the PDB and made available while the structure is still

"on hold," and sequences from TargetDB (Chen et al., 2004), for which a structure has been solved by another structural genomics center. We also included sequences of BSGC targets that have progressed to the "Traceable Map" stage, as this usually indicates the structure will soon be completed. The database was updated prior to each target selection round.

3. snr: the non-redundant set of protein sequences from Swiss-Prot (Boechmann et al., 2003). We included all sequences in the swissprot, trembl, and trembl_new files (downloaded 30 July 2001 for round 2, 30 November 2001 for round 3, 21 October 2002 for rounds 4-5 and 7-8, and 26 February 2004 for round 6) from Swiss-Prot, which had been filtered with the SEG (Wootton, 1994) and PFILT (Jones and Swindells, 2002) programs using default options. The filtering was done to reduce the chance of profile corruption (Schaffer et al., 2001), which can lead to inaccurate results.
4. Available genomes: NCBI database of proteins from sequenced bacterial and archaeal genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). Targets were only chosen from genomes for which the BSGC had access to purified genomic DNA. These species are listed on the BSGC website (http://www.strgen.org).

Identification of known structures: At the beginning of each round of target selection, all *MP* proteins and their homologs were considered potential targets. These were then removed from consideration if they were detectably homologous to other proteins of known structure.

1. In each automated target selection round, sequences of all *MP* ORFs were compared to the knownstr database using several sequence comparison tools such as PSI-BLAST (Altschul et al., 1997). PSI-BLAST position-specific scoring matrices (PSSMs) were constructed for each *MP* ORF using 10 rounds of searching our "snr" database with a matrix inclusion threshold E-value of 10^{-2} in most rounds.
2. The PSSMs were used to search the knownstr database, and any hits with an E-value of 10^{-1} or below were eliminated from consideration as targets. This significance threshold was chosen to increase the likelihood of detecting more remote homologs, even though it had some risk of false positives being removed from the target list.
3. After the second round for target selection, the matrix inclusion threshold was increased in order to increase the possibility of identifying remote homologs, at the risk of a higher rate of corrupted PSSMs.

4. Because of the latter possibility, we also used BLAST (Altschul et al., 1990) and Pfam (Bateman et al., 2004) in target selection rounds 3-6. All *MP* ORFs with a BLAST hit against knownstr with an E-value of 10^{-1} or below were eliminated from consideration as targets, in addition to those already eliminated by PSI-BLAST.
5. Pfam was also used to detect known structures. The HMMER tool (Eddy, 1998) was used to compare the Pfam_ls library of hidden Markov models to both the knownstr database and the database of *MP* ORFs, using the family-specific "trusted cutoff" score as a cutoff for assigning significance. We eliminated from consideration all ORFs that had a significant hit to a Pfam family that had also matched at least one known structure.

Identifying *MP* targets predicted to be less tractable for high-throughput study:

1. As the next step in each target selection round, we eliminated *MP* proteins and domains that were likely to be predictably less tractable for high-throughput study. These included proteins with regions of amino acids predicted to be in transmembrane segments, coiled coils, and regions of low complexity. The predictions were made by the SEG program (version dated 24 May 2000) for proteins with low complexity regions spanning more than 20% of the protein lengths, the CCP program (written by J Kuzio at NCBI, version dated 14 June 1998), using the algorithm of Lupas (Lupas, 1996), for proteins with coiled coil regions, and two programs to identify transmembrane regions, TMHMM 2.0a (Krogh et al., 2001) and PHDhtm (Rost et al., 1995) version 2.1 (October 1998). Any transmembrane region predicted by either program eliminated an *MP* ORF from consideration as a target in rounds 2-5.
2. We also eliminated potential targets that were long and therefore likely to be challenging; in earlier rounds (round 1-2) of target selection, the length cutoff was 400 amino acids, and in later rounds (round 3-8) it was increased to 700 amino acids.
3. Finally, we excluded proteins annotated as ribosomal components, as these were expected to be unlikely to be stable in the absence of binding partners.

Identifying homologues of *MP* proteins as targets: In addition to the *MP* proteins themselves, homologous proteins from other prokaryotes were also chosen as targets. Each *MP* protein (or predicted domain in round 6) that passed through the above filters was used to search the database of available genomes using PSI-BLAST. PSI-BLAST PSSMs were constructed for each *MPe* ORF using 10 rounds of searching the nonredundant sequence database "snr" (as described above) with default parameters; the PSSMs were then used to search the database of genomes. BLAST version 2.2.4 was also used (with default parameters) in rounds 4-8 to search the genome database. All proteins identified by BLAST or PSI-BLAST with E-values more significant than 10^{-4}, with the region of local similarity covering at least 50 residues, were considered as possible targets.

Other factors considered: Potential targets from *MP* were always selected if they passed an additional screen to ensure they could be expressed in the *Escherichia coli* expression system used at the BSGC. *MP* and other related mollicutes such as *Ureaplasma urealyticum* can use UGA codons to encode the amino acid tryptophan, whereas UGA is a stop codon in *E. coli*. Thus, cloned *MP* proteins with this codon would express truncated proteins in *E. coli*. In cases where a UGA codon was within about 30 bases of either end of the gene, it could easily be mutated to a UGG codon during cloning, using mutating PCR primers. Other UGA codons, called internal UGA codons, could only be mutated in a more difficult multi-step cloning procedure.

When there were too many homologous targets, high priority was given to targets from thermophiles and halophiles, as these were expected to be experimentally more tractable, for example being partially purified by heating the *E. coli* lysate.

Target De-selection: The BSGC only seeks to solve structures for protein domains for which the structure cannot be reliably predicted via bioinformatic methods. We therefore de-select and stop work on targets whose structures of similar proteins have been solved by other groups. Most deselection analysis steps are automated. However, the final decision on whether to stop work on a target is performed manually, to decrease work lost due to potential false positives. This automated analysis and manual review are both performed weekly. More details of the rationale behind this two step approach are given elsewhere (Chandonia et al., 2006).

IV. Protein Production

For our purpose, *E. coli* recombinant expression systems are the best option in terms of economy and ease of protein production. Prokaryotic cell-free protein synthesis has also been used occasionally.

Cloning: For the past two years, BSGC has used an in-house version of the Ligation Independent Cloning (LIC) methodology (Aslanidis & de Jong, 1990). This LIC system provides both efficient high-throughput cloning and flexibility in fusion construction. The LIC method relies on common linker sequences to anneal and join the target segments to the vector. A Tobacco Etch Virus (TEV) cleavage site allows cleavage of the fusion tag. The fusions (MBP, GST, TRX, NusA) are utilized primarily for enhancing soluble expression. In addition to the N-terminal His$_6$ tag, we have a PCR-based method of preparing targets containing a C-terminal His$_6$ tag that can be used in cases where the N-terminal His$_6$ tag is ineffective. The simplicity of the present LIC cloning scheme allows for most of the experimental steps to be performed robotically in groups of 96 targets. The following steps are currently automated on the Biomek 2000 (Beckman Coulter, Inc., Fullerton, CA) robot: PCR reaction setup and cleanup, PCR product analysis by E-gel 96 (Invitrogen, Inc., Carlsbad, CA), LIC reaction and transformation, Mini-expression setup, Clone preservation in agar stab, and Plasmid preparation.

Small-Scale Expression: After transformation into the expression host, two colonies are selected and grown in an auto-inducing medium (Grabski et al., 2003). Cells are grown in a 96 deep-well plate overnight, spun down, resuspended, and sonicated using the Misonix 3000 sonicator (Misonix, Farmingdale, NY). The lysate is spun, and both soluble and insoluble fractions are run on SDS/PAGE. Presently, **all steps**, from PCR reaction for 96 targets to analysis of level of expression of the targets, are automated and can be achieved in 5 days (Nguyen et al., 2004).

Large-Scale Preparation of Cell Paste: Previously we had used Luria broth with isopropyl-β-D-1-thiogalactopyranoside (IPTG) induction. This required the manual addition of IPTG. For the last 2 years we have used an auto-inducing medium (developed by Dr. William Studier from Brookhaven National Lab) that induces expression by balancing the levels of glucose and lactose as carbon sources. This formulation spontaneously induces high levels of target protein without the need to monitor growth and increases the soluble expression of target proteins. Two types of auto-inducing media are used: 1. ZYP: native medium, 2. PASM: medium for labeling the target protein with Seleno-methionine.

Protein Purification: Parallel purification is performed on three AKTA Explorer work stations (GE Healthcare, Piscataway, NJ). We have recently changed our protocol to the following: 5 targets are sequentially purified through three columns in an automated way using the AKTA Explorer with 3D Kit (software necessary for programming these steps). The three columns being used are: HisTrap metal-chelating—Desalting column—HiTrap Q/S HP 5 ml column (GE Healthcare, Piscataway, NJ). This method takes 10 hr to complete.

Quality Control Assessments:

All purified proteins undergo quality control steps 1–5 as listed in Table 1. 1-D NMR is performed on proteins smaller than 45 kDa that did not crystallize.

Table 1. Protein Characterization.

Parameters	**Method**
1. Purity	SDS/PAGE stained with Coomassie Brilliant Blue R
2. Monodispersity	Dynamic light scattering (DynaPro 99, Wyatt Technology Corp., Santa Barbara, CA)
3. Aggregation state	Native gel (Phast system, GE Healthcare, Piscataway, NJ) Analytical size exclusion chromatography (G4000SWxl, Tosohaas Corp., Montgomeryville, PA)
4. Molecular weight	Mass spectrometry (MALDI-TOF, Voyager DE, Applied Biosystems, Foster City, CA)
5. Bound elements	ICP-MS (University of Georgia, Athens, Ga)
6. Functionality	Panel of enzymatic assays (In collaboration with Dr. A. Yakunin, University of Toronto, Toronto, Canada)
7. 1-D NMR	Bruker DRX 500 NMR spectrometer using an 11 (one-one) pulse sequence (Dr. D. Wemmer, University of California, Berkeley, CA)

Protein production summary: As mentioned in the Introduction, BSGC is unique in that we have chosen as our target two minimal organisms (*MP* and *MG*) with the smallest genome size. Our targets have no sequence homology to those of known structures. Even with a small starting pools of targets, a multi-path approach eventually allowed us to produce most of our targets. For the minimal organism MP with 677 full-length predicted proteins, after filtering out the proteins that are structural homologs of known structures, those containing transmembrane domains, coiled coils, low complexity regions, and multiple UGA codons, there remained **82** full-length target genes. Up to 10 homologs for each gene from other organisms were added to make a total of 386 targets. Out of 386 targets, 318 were successfully cloned and 261 clones gave good expression. From those, 191 proteins were purified in good quality and amounts suitable for crystallization screening.

V. Technical Development for Challenging Proteins

Heat Shock and High Salt Growth: Over-expression of many heterologous proteins results in production of refractive bodies, also known as inclusion bodies (IB). The level of these insoluble proteins can sometimes be reduced by lowering the growth temperature upon induction; changing the media composition; expressing the protein as a fusion with MBP, GST, thioredoxin, or NusA (Sachdev and Chirgwin, 1998; Harrison, 2000); and inducing the expression of chaperones. Other approaches for reducing IB production are salt and heat-stress which induce complementing defense mechanisms in bacterial cells, including intracellular accumulation of osmolytes or synthesis of heat-shock proteins, respectively (Kempf and Bremer, 1998; Bukau and Horwich, 1998). Simple heat shock before induction is known to enhance the solubility of some recombinant proteins produced in *E. coli* (Chen et al., 2002). Some osmolytes behave as "chemical chaperones" by promoting the correct folding of unfolded proteins *in vitro* and in the cell (Samuel et al., 2000; Yang et al., 1999; Voziyan and Fisher, 2000; Diamant et al., 2001). We have combined these two elements, heat shock and high salt media, to increase the fraction of soluble protein produced from targets.

We have tested a protocol that combines heat shock and high salt growth (Oganesyan et al., 2006). The cells were grown in the presence of 0.5 M NaCl and incubated at 47°C at the beginning of induction with IPTG for 20 minutes. The temperature was then decreased to 20°C for overnight growth. These cells expressed only soluble target, although the total level of expression was 10 fold lower than when grown under "normal" conditions. This soluble sample was crystallized, and its structure was solved (Das et al., 2004).

On-Column Refolding: Inclusion body formation, as mentioned above, can be minimized or avoided by applying complex efforts to enhance production of soluble protein. On the other hand, protein production from inclusion bodies has a number of merits. They are: (1) produced in high yields, even those that are toxic for bacterial cells; (2) generally protected from proteolytic degradation; (3) and easily purified and solubilized. The main challenge is to convert inclusion bodies to properly folded, biologically active proteins. We have developed an on-column chemical refolding method (Oganesyan et al., 2004), for insoluble His-tagged proteins expressed in *E. coli*, partly based on the method described by Rozema and Gellman (1996). IBs solubilized in urea are first bound to a metal-chelating affinity column and exposed to a detergent wash to prevent misfolding. This is followed by a β-cyclodextrin wash that removes the detergent and promotes correct folding

(Daugherty et al., 1998). The target protein is eluted with Imidazole, and then goes through further purification steps—IEX and/or SEC—before evaluation by Dynamic Light Scattering (DLS). As an example, ten of the PSI-1 targets from BSGC that expressed insoluble protein were purified using this method. Three of the ten targets could not be refolded, but we obtained 30–100% refolding from the other seven. All refolded proteins were subjected to DLS analysis, and five out of seven refolded proteins were monodisperse. Six of the seven refolded proteins were able to produce crystals of varying qualities.

Optimum Solubility (OS) Screen: For structural studies, the first step after a protein is purified is to concentrate it in its purification buffer to a concentration suitable for crystallization or NMR studies. In about 25% of the cases, this step fails because the protein aggregates and precipitates; this adverse phenomenon is totally unpredictable. Inspired by a screen for NMR studies (LePre & Moore, 1999), we have developed a screening method (Jancarik et al., 2004) in which we test a panel of buffers and many additives to obtain the most homogeneous and monodisperse solution for each protein that usually aggregates and cannot be concentrated prior to setting up crystallization screens.

A panel of 24 buffers is tested using the hanging-drop method and vapor diffusion equilibrium. After monitoring precipitation, the conditions leading to clear drops are selected for DLS characterization. For this part of the screen, only 24 μl of protein (with concentration ranging from 3 mg/ml to 10 mg/ml) are required. If the DLS results are not optimal, a series of additives are tested in the presence of the best buffer selected from the initial screen, and again DLS is used to determine the best condition. The OS screen has been performed on 14 samples of cytoplasmic proteins that had aggregated as measured by DLS and had precipitated upon concentration or could not be concentrated. The OS screen indicated that out of the 14 protein samples, the DLS of 11 of them could be improved in different buffers, and, in some cases, an additive further improved DLS. Nine of these proteins could subsequently be crystallized.

VI. Structure Determination

The overall flow of the process for determining three-dimensional (3D) structures of proteins is well established. Although much of the science involved is understood, and many of the key techniques are well developed, the underlying science is not well understood in some steps, and so the outcome appears almost stochastic and unpredictable. Thus from an engineering and automation point of view, the component steps in the process from purified protein to 3D structure can be divided into two categories: (1) the steps that are automatable and can be operated in a high-throughput mode, and (2) the steps that can only be processed in a multi-path approach by screening a large number of conditions, factors, and paths to increase the probability of success for such steps. The success rate for the single-path approach from purified protein to unique structure is, on average, about **9%**. In the PSI-1 pilot stage, despite the limited manual multi-path approach, we have been able to increase the success rate to ~**27%** (the corresponding success rate for **clone to structure** is about 5% and 16%, respectively) with additional multi-path steps and automation. The overall pipeline at BSGC is schematically shown below.

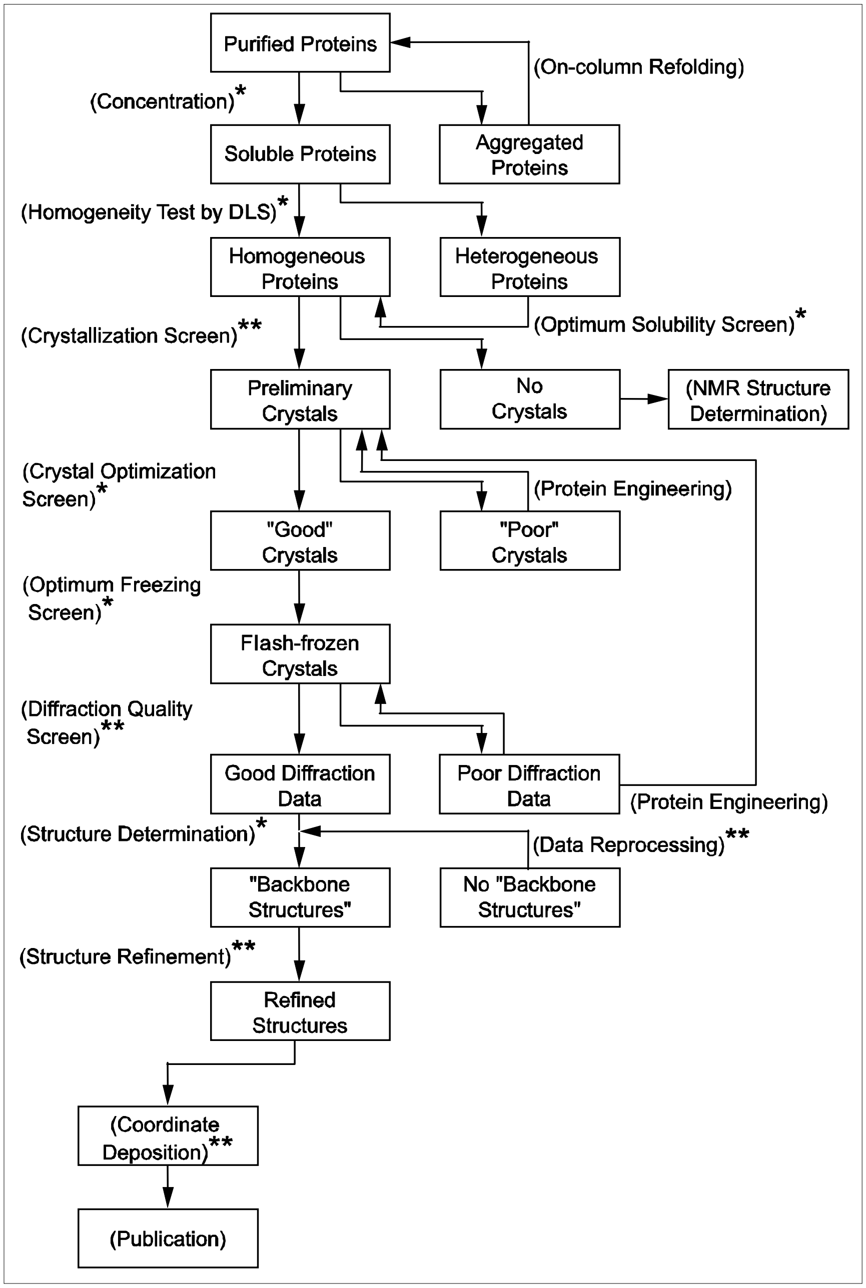

Fig. 3. Multi-path flow diagram for the process from pure proteins to 3D structures. The process for each step is in parenthesis. The automated steps are marked by **; steps that are partially automated or steps for which screens have been developed but haven't been automated are marked by *.

Crystallization: The science of protein crystallization is not well understood. Currently, the most successful and practical method for finding protein crystallization conditions is to screen a large number of conditions through our sparse matrix crystallization screening method (Jancarik & Kim, 1991) and its commercially available variations (e.g., from Hampton Research, Aliso Viejo, CA). During this process, we use the Hydra Plus One (Matrix Technologies Corp., Hudson, NH) and the Phoenix Liquid Handling System (Art Robbins Instruments, Sunnyvale, CA) crystallization robots with 96 well plates. We routinely screen 4 x 96 crystallization conditions at two temperatures. Once one or more promising crystal hits are found, we optimize the hit conditions using a protocol we have developed to fine-tune the conditions to obtain good diffraction quality crystals. As a result of the on-column refolding step and the optimum solubility screen described earlier, our success rate for the purified-protein-to-diffraction-quality-crystal process is about 27%.

Diffraction Data Collection: Many of the steps in diffraction data collection at Berkeley Lab's Advanced Light Source are hardware and software assisted. They include the robotized automatic mounting of frozen crystals, point-and-click crystal centering, and the capability to screen frozen crystals to search for well-diffracting crystals.

Structure Solution: Once experimental data are collected, high-throughput methods are applied to solve and complete a structure. We routinely use software developed for structural genomics efforts, such as HySS for substructure determination (Grosse-Kunstleve & Adams, 2003). This software is part of the PHENIX package. The high level of automation HySS provides makes it possible to determine a substructure at the beamline immediately after data have been collected and processed. Once the anomalous substructure has been located, phase calculation and substructure refinement are performed, using SOLVE (Terwilliger and Berendzen, 1999), MLPHARE (CCP4, 1994), or CNS (Brunger et al., 1998) as dictated by data quality. In more challenging cases the SHARP (de La Fortelle and Bricogne, 1997) program is used. The pipeline is shown below.

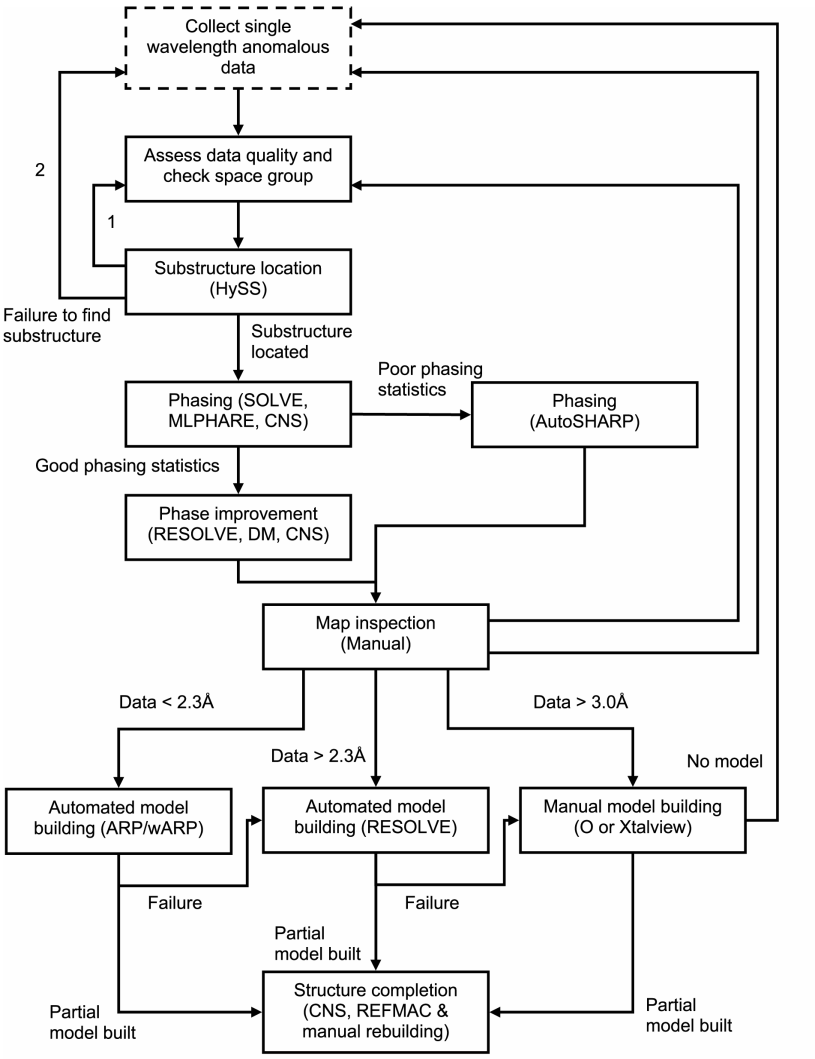
Collect single wavelength anomalous data
Assess data quality and check space group
2
1
Substructure location (HySS)
Failure to find substructure
Substructure located
Phasing (SOLVE, MLPHARE, CNS)
Poor phasing statistics
Phasing (AutoSHARP)
Good phasing statistics
Phase improvement (RESOLVE, DM, CNS)
Map inspection (Manual)
Data < 2.3Å
Data > 3.0Å
Data > 2.3Å
No model
Automated model building (ARP/wARP)
Automated model building (RESOLVE)
Manual model building (O or Xtalview)
Failure
Failure
Partial model built
Structure completion (CNS, REFMAC & manual rebuilding)
Partial model built
Partial model built

Fig. 4. Flowchart for structure solution, model building, and structure completion used by BSGC in the PSI pilot phase.

The results of phasing are continued into phase improvement by density modification, using CNS (Brunger et al., 1998), DM (Cowtan, 1999), and RESOLVE (Terwilliger, 2000; Terwilliger, 2002). Visual inspection of the electron density map is used to determine whether more experimental data should be collected. For model building we use automatic software when possible. If data extend to 2.2 Å, the ARP/warp (Perrakis, Morris & Lamzin, 1999) suite is used, and in a favorable case 90% of the model is built. With lower resolution data (between 3 and 2.2 Å) the RESOLVE software is used to build an initial model, typically at least 50% of the main chain is built. The model is then used as a basis for manual model completion. In cases of poor data quality or resolutions below 3.0 A, a manual model building is used. Structure refinement and model completion makes use of the standard refinement tools: CNS and REFMAC (Murshudov, Vagin & Dodson, 1997), automated water assignment, and manual rebuilding if necessary.

VII. Summary of BSGC Throughput during the PSI-1 Pilot Phase

For the minimal organism *MP* with 677 full-length predicted proteins, after filtering out the proteins that are structural homologs of known structures, those containing transmembrane domains, coiled coils, low complexity regions, and multiple UGA codons, there remained 82 full-length target genes. Up to 10 homologs for each gene from other organisms were added to make a total of 386 targets. Of those, 318 were successfully cloned (not counting clones of domains of full length proteins). Our overall success rates are shown in Table 2.

Table 2. BSGC success rate for full length target proteins during the PSI-1 pilot phase.

Full length genes cloned	Soluble expression	Purified proteins	Crystallized	Structures solved
318	261	191	104	93

BSGC Solved Structures: Almost all of BSGC targets are "unique" in that the majority of the targets have no sequence homologs among proteins of known structures. Thus we have had a high rate of discovering many new protein folds. A number of these also revealed unexpected bound ligands suggesting their possible biochemical functions, and others have unusual oligomeric structures not predicted by genetic or biochemical methods. Thus, the majority of BSGC structures belong to one of four categories: (1) hypothetical proteins with novel folds, (2) proteins with novel folds that suggest their molecular functions, (3) proteins with "unique" sequences that reveal novel folds, and (4) hypothetical proteins with known folds ("remote homologs"). The protein structures in categories 2 and 4 can infer possible molecular functions (Kim et al, 2004).

IX. Structural proteome of a minimal organism

When the genomic sequence of the first organism was completed, development of computational methods to analyze the sequenced genes became the key for extracting valuable new information, most of which was totally unpredicted and unexpected. The critical importance of the computational methods became even more evident as more genomic sequences became available. As was the case with sequence genomics, the development of computational methods for analysis of the 3D structures of proteins is going to be the key to mining valuable information from the 3D structures of proteins obtained from PSI and other sources. Toward this objective, we have developed a computational process to represent all unique protein structures in a multidimensional space based on structural similarities and in 3-D space for approximate visual representation of the multidimensional structural space.

The Protein Structure Space Mapping: The PSI objective of near-complete coverage of protein structure space needs a representation method of the space. It has been shown recently (Hou et al., 2003, 2004) that the protein structure space can be "mapped" in three-dimensions, in which all the known and newly determined protein structures are distributed in a highly organized way. Furthermore, the demographic distribution of the protein structures in the map is understandable from the viewpoints of protein architectural features and protein fold evolution. Thus, this representation of the protein structure space provides a *unified platform* on which all the protein structures of the PSI, as well as others, can be mapped, and various structural information, functional information, and evolutionary information can be mapped and mined computationally, once such computational tools are developed.

Mapping of the Protein Structure Space: One of the major objectives of PSI is to obtain a broad coverage of the protein structure space. To conceptualize the space, and derive new information from the demographic distribution of protein structures in the space, it is useful to define it. Calculating all pair-wise structural similarity for all non-redundant protein structures (~ 2000) in PDB, and converting them to structural dissimilarity scores, we were able to map the protein structure space in three-dimensional space. To accomplish this, we used the mathematical method known as multidimensional scaling (Hou et al., 2003, 2004). In the structural space, each point represents a unique protein structure family. In this space, each point is located in the space that best fits all pair-wise distances between the point and all the rest. Two points are close to each other when their structures are similar. The following observations were made:

1. The protein structure space is sparsely populated, and all protein structures are confined to four elongated regions, each characterized by particular architectural features of proteins. This observation strongly suggests that evolution of proteins may have been strongly restricted by the requirement of architectural stability of proteins.
2. Short and poorly structured proteins are mapped near the "origin," and the size of proteins and the extent of secondary structure, or super-secondary structure, elements generally increase along each feature axis as indicated in Fig. V.6. This suggests that these trends may be related to protein-fold evolution.

3. The three feature axes and the three coordinate axes provide a completely general and objective way of describing and classifying protein structures, thus providing a new demographic addressing a system similar to library cataloguing. Furthermore, these structure features are easily computed from protein structure information without solving any structural alignment optimization algorithms, so they may serve as basic feature vectors for a fast, generalized, and automatic protein structure classifier.
4. All new structures from the PSI program and others map approximately within the "envelope" defined by the structural space originally found using ~2000 non-redundant structures of PDB, suggesting that the "protein structure universe" is finite.

This type of representation of protein structure space provides a unified platform on which one can map all the PSI structures and others, in order to globally visualize the structural relationship among them, to identify the regions of different structural population densities for suggesting additional new structures needed, and to infer possible protein-fold evolution. Furthermore, all biochemical and biophysical functions can be mapped on the space to obtain a global view of the molecular function-structure classification and fold/function relationship on a global level.

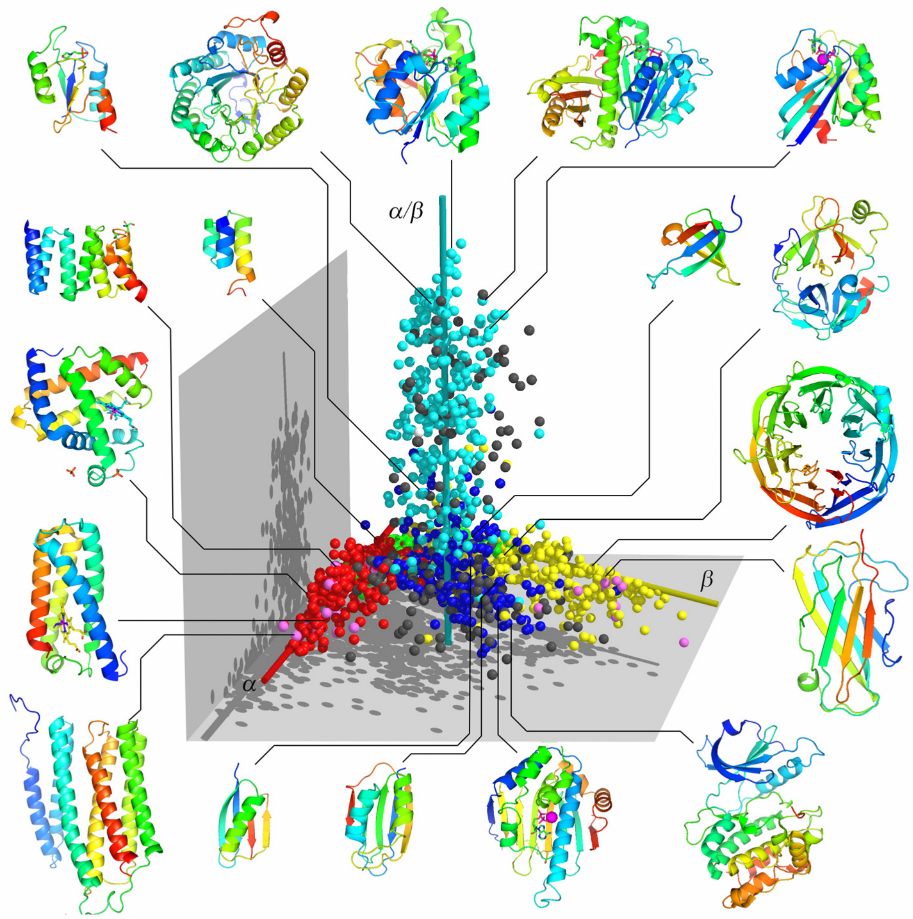

Fig. 5 Global representation of protein fold space. All together, 1898 unique protein structure families are represented by spheres in the three-dimensional space. α, β and α/β class structures follow three elongated directions denoted by α, β and α/β feature axes (Hou et al., 2004). Class designation for each structure family according to the SCOP (Murzin et al., 1995) database is indicated by red for α-class, yellow for β-class, blue for α+β class, cyan for α/β class, pink for membrane proteins, and black for multidomain proteins. In most cases, SCOP classification approximately agrees with the demographics of the protein fold space. Some sample structures are shown.

Structural families found in the minimal organism, *MP:* The unique structural families represented by all the soluble MG/MP proteins and their homologs are mapped on the protein structure universe. As expected they are located within the envelope defined earlier by the 1898 non-redundant structures of PDB. There appears to be a paucity in β-class proteins in these minimal organisms.

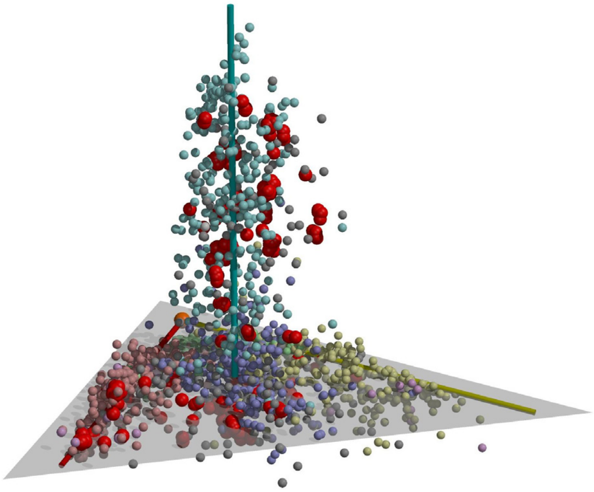

Fig.6. Protein structure families determined at BSGC (red) mapped on the protein structure universe. Most of the BSGC structures had no sequence homologues in PDB structure database. About one half of them had new folds and occupy empty spaces in the protein structure universe, and the other half turned out to be "remote homologues" of structures of known folds and occupy the same or very close to pre-occupied locations.

Structural coverage of a minimal oganism, *MG:* At the start of PSI-1, about 2/3 of the MG proteins had no structural information, of which the majority (about 43% of total) were predicted to be soluble proteins (see Fig. 6a). At the end of

PSI-1, we now have structural information for over 90% of the soluble proteins of this minimal organism (Fig. 6b) (Chandonia and Kim, 2006).

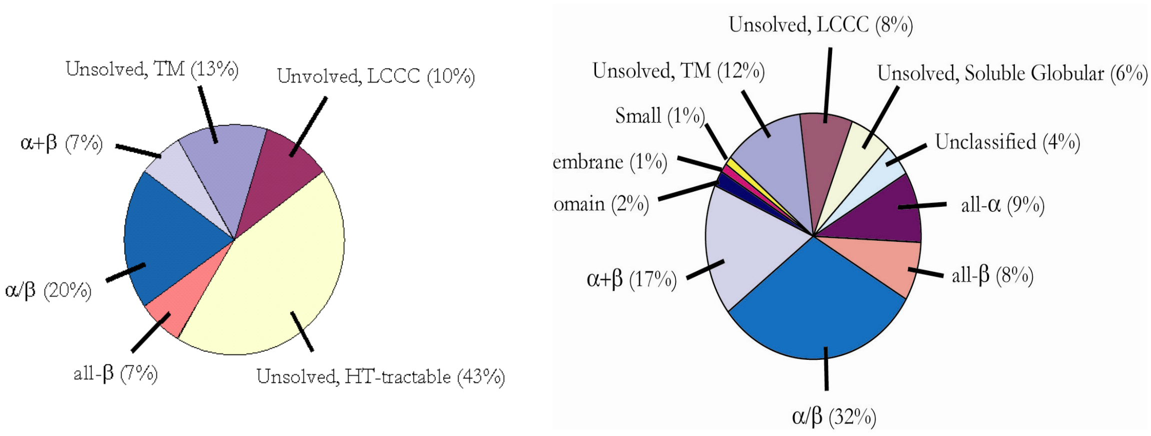

Fig. 7 (Left) At the start of PSI-1, 3-D fold information was available for 34% of the proteins in *Mycoplasma genitalim*, the rest of the proteins belonged to membrane proteins (13%), low complexity proteins (10%) and soluble proteins of unknown 3-D folds (43%). (Right) By the end of PSI-1, 3-D fold information was available for over 90% of soluble proteins in this minimal organism.

Further analysis of this and other structural proteomes of small prokaryotes reveals an interesting conservation pattern for protein fold for proteins of particular functional categories, details of which were described recently (Chandonia and Kim, 2006).

Acknowledgements

This work is supported by grants from the NIH (1-P50-GM62412 and 1-R01-GM073109). We are grateful to a large number of colleagues who participated in various aspects of BSGC's PSI-1 program, such as high throughput cloning and expression, large-scale production and characterization of proteins, crystallization and structure determination, and structural space mapping, with the guidance of Drs. S. Brenner, D. Wemmer, T. Earnest, D. McKay, C. Hutchison Jr., J. T. Hou, and I.-G. Choi

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res* **25,** 3389-402.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J Mol Biol* **215,** 403-10.

Aslanidis, C., & De Jong, P. J. (1990). Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res.* **20,** 6069-74.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004) *Nucleic Acids Res* **32 Database issue,** D138-41.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Res* **28,** 235-42.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) *Nucleic Acids Res* **31,** 365-70.

Brunger, A.T., Adams, P. D., Clore, G. M, DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., & Warren, G.L. (1998). Crystallography & NMR system: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallographica.* **D54**, 905-921.

Bukau, B., & Horwich, A. L. (1998). The Hsp70 and Hsp60 chaperone machines.
Cell. **92**, 351-366.

Chandonia, J. M., Kim, S. H., and Brenner, S. E. (2006) *Proteins* **62,** 356-70.

Chandonia, J. M., and Kim, S. H. (2006) *BMC Struct Biol***,** in press.*****

Chen, L., Oughtred, R., Berman, H. M., and Westbrook, J. (2004) *Bioinformatics.*

Chen, J., Acton, T. B,, Basu, S. K., Montelione, G.T., & Inouye, M. (2002). Enhancement of the solubility of proteins overexpressed in *Escherichia coli* by heat shock. *J Mol Microbiol Biotech.* **4**, 519-24.

Cowtan, K. (1999). Error estimation and bias correction in phase-improvement calculations. *Acta Cryst.* **D55**, 1555-1567.

Das, D., Oganesyan, N., Yokota, H., Pufan, R., Kim, R., and Kim, S.-H. (2004). Crystal structure of the conserved hypothetical protein MPN330 (GI: 1674200) from *Mycoplasma pneumoniae*. *PROTEINS: Struc. Func. Bioinf.* (In Press).

Daugherty, D. L., Rozema, D., Hanson, P. E., & Gellman, S. H. (1998). Artificial Chaperone-assisted Refolding of Citrate Synthase. *J. Biol Chem.* **273**, 33961-33971.

De La Fortelle, E. & Bricogne, G. (1997). Maximum-Likelihood Heavy-Atom Parameter Refinement in the MIR and MAD Methods. *Methods Enzymol.* **276**, 472-494.

Diamant, S., Eliahu, N., Rosenthal, D., & Goloubinoff, P. (2001). Chemical chaperones regulate molecular chaperones in vitro and in cells under combined salt and heat stresses. *J Biol Chem.* **276**, 39586-91.

Eddy, S. R. (1998) *Bioinformatics* **14,** 755-63.

Grabski, A., Mehler, M. & Drott, D. (2003). Unattended high-density cell growth and induction of protein expression with the Overnight Express Autoinduction System. *Innovations.* **17**, 3-6.

Grosse-Kunstleve, R.W. and Adams, P.D. (2003). Substructure search procedures for macromolecular structures. *Acta Cryst.* **D59**, 1966-1973.

Harrison, S. C. (2004). Whither structural biology? *Nat. Struct. Mol. Biol.* ***11***, 12-15.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C., and Herrmann, R. (1996) *Nucleic Acids Res* **24,** 4420-49.

Hou, J., Sims, G. E., Zhang, C. & Kim, S. -H. (2003). A global representation of the protein fold space. *Proc. Nat. Acad. Sci.* **100**, 2386-2390.

Jancarik, J. and Kim, S. H. (1991). Sparse matrix sampling: a screening method for crystallization of proteins. *J. Appl. Cryst.* **2**, 409-411.

Jancarik, J., Pufan, R., Hong, C., Kim, R., Kim, S. –H. (2004) Optimum Solubility (OS) Screening: an efficient method to optimize buffer conditions for homogeneity and crystallization of proteins. Acta Cryst. **D60:** 1670-1673.

Jones, D. T., and Swindells, M. B. (2002) *Trends Biochem Sci* **27,** 161-4.

Kempf, B. & Bremer, E. (1998). Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments. *Arch. Microbiol.* **170**, 319-330.

Kim, S. H., Shin, D. H., Choi, I. G., Schulze-Gahmen, U., Chen, S., Kim, R. (2003) Structure-based functional inference in structural genomics. J. Struct. Funct. Genomics **4**: 129-135.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001) *J Mol Biol* **305,** 567-80.

Lupas, A. (1996) *Methods Enzymol* **266,** 513-25.

Murshudov, G. N., Vagin, A. A., & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst.* **D53**, 240-255.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) *J Mol Biol* **247,** 536-40.

Nguyen, H., Martinez, B., Oganesyan, N., Kim, R. (2004) An automated small-scale protein expression and purification screening provides beneficial information for protein production. J. Struct. Funct. Genomics **5**: 23-27.

Oganesyan, N., Ankoudinova, I., Kim, S.-H., and Kim, R. (2006) Effect of Osmotic Stress and Heat Shock in Recombinant Protei Overexpression and Crystallization Protein Expression and Purification (In Press).

Oganesyan, N., Kim, S. –H., Kim, R. (2004) On-column Chemical Refolding of Proteins. PharmaGenomics **4**: 22-26.

Perrakis, A., Morris, R., & Lamzin, V. S. (1999). Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.* **6**(5), 458-63.

Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1995) *Protein Sci* **4,** 521-33.

Rozema, D. & Gellman, S.H. (1996). Artificial chaperone-assisted refolding of denatured-renatured lysozyme : modulation of the competition between renaturation and aggregation. *Biochemistry* **35**, 15760-15771.

Sachdev, D. & Chirgwin, J. M. (1998). Solubility of proteins isolated from inclusion bodies is enhanced by fusion to maltose-binding protein or thioredoxin. *Protein Express. Purif.* **12**, 122-132.

Samuel D., Kumar, T. K, Ganesh, G., Jayaraman, G., Yang, P. W., Chang, M. M., Trivedi, V. D., Wang, S. L., Hwang, K. C. & Chang, D. K. & Yu, C. (2000). Proline inhibits aggregation during protein refolding. *Protein Sci.* **9**, 344-352.

Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., and Altschul, S. F. (2001) *Nucleic Acids Res* **29,** 2994-3005.

Terwilliger, T. C. (2004): SOLVE and RESOLVE: automated structure solution, density modification and model building. *J Synchrotron Radiat* **11**, 49-52.

Terwilliger, T. C. & Berendzen, J. (1999). Automated MAD and MIR structure solution. *Acta Crystallogr D Biol Crystallogr.* **55**(Pt 4), 849-61.

Voziyan, P. A. & Fisher, M. T. (2000). Chaperonin-assisted folding of glutamine synthetase under nonpermissive conditions: off-pathway aggregation propensity does not determine the co-chaperonin requirement. *Protein Sci.* **9**, 2405-2412.

Wootton, J. C. (1994) *Comput Chem* **18,** 269-85.

Yang, D. S., Yip, C. M., Huang, T. H, Chakrabartty, A. & Fraser, P. E. (1999). Manipulating the amyloid-beta aggregation pathway with chemical chaperones. *J. Biol. Chem.* **274**, 32970-32974.

Busso, D., Kim, R**.,** Kim, S.-H. (2004) Using an *Escherichia coli* cell-free extract to screen for soluble expression of recombinant proteins. J. Struct. Funct. Genomics **5**:69-74.