Section V

Structural Proteomics Initiatives Overviews

Chapter 32

Structural Genomics of Minimal Organisms: Pipeline and Results

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

The initial objective of the Berkeley Structural Genomics Center was to obtain a near complete three-dimensional (3D) structural information of all soluble proteins of two minimal organisms, closely related pathogens *Mycoplasma genitalium* and *M. pneumoniae*. The former has fewer than 500 genes and the latter has fewer than 700 genes. A semiautomated structural genomics pipeline was set up from target selection, cloning, expression, purification, and ultimately structural determination. At the time of this writing, structural information of more than 90% of all soluble proteins of *M. genitalium* is available. This chapter summarizes the approaches taken by the authors' center.

1. Introduction

1.1. Mission

The Protein Structure Initiative (PSI) of US National Institutes of Health (NIH) aims to obtain structural information on all proteins derivable from their DNA sequences (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) identify critical areas and steps for further development to achieve a high throughput operation; and (3) 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.

1.2. 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 *Mycoplasma genitalium* [MG], with ~700 and ~500 genes, respectively). This objective is accomplished for BG at the greater than 90% level.

From: *Methods in Molecular Biology, Vol. 426: Structural Proteomics: High-throughput Methods*
Edited by: B. Kobe, M. Guss and T. Huber © Humana Press Inc., Totowa, NJ

1.3. Pipeline

To achieve this, the authors have developed methods and protocols to automate or parallelize many processes from cloning the target genes to structure determination. Overall pipeline schemes for the single-path approach used in the initial 2-year period and the multiple-path approach used for the rest of the PSI-1 period are shown in Fig. 32.1.

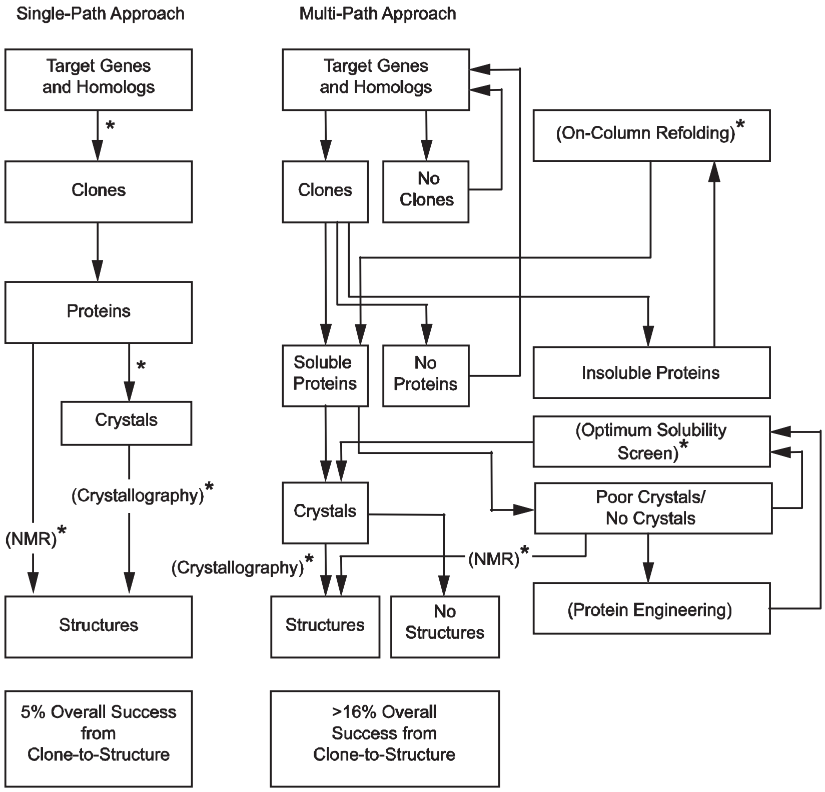

Fig. 32.1 Single-path approach vs. multipath approach for soluble proteins. A large number of target genes and their homologues (coding for proteins with no sequence homologies to the proteins of known structures) were selected (see Target Selection), and the majority of them could be cloned. Of these, some were overexpressed as proteins in soluble form, protein aggregates, or insoluble inclusion bodies. In the single-path mode (low-hanging fruits), 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 multipath mode, those clones not expressing, or underexpressing could be recloned with different constructs and/or into different vectors to obtain additional overexpressing clones; proteins that aggregate or could not be concentrated underwent optimum solubility screening (see the following) to find optimum conditions in which they were soluble and homogeneous. Those proteins that were insoluble underwent an on-column refolding process (see the following). Proteins that were chemically and conformationally homogeneous were used for NMR or crystallization studies. BSGC experience shows that for this multipath approach the overall success rate increased to about 16%. (*Processes that are automated or semiautomated.)

2. Metrics and Lessons Learned

Based on BSGC's results during the PSI-1 period, the metrics and lessons required for a structural genomics approach to a large-scale structure determination effort were learned; 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 in the following:

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 in which the underlying science and technologies are well understood (thus, *automatable* by instrumentation or programming); and (2) those in which 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 *multivariable screenings*.
2. The *single-path approach* (see Fig. 32.1), 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 similarity to those of known structures in the Protein Data Bank (PDB) *(1)*.
3. The *multipath approach* (see Fig. 32.1), in which feedback loops and multifactor 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 new folds, and the remaining half are "remote homologues," structures similar to known structures without sequence similarity (similar structure without sequence similarity) of known folds.
5. Approximately two thirds of the structures of "hypothetical proteins" (proteins that have no sequence homologues 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 3D space based on pairwise structural similarities *(2,3)*; thus providing a platform for representing all protein structures, "the protein structure universe."

3. Selection of Target Proteins for High Throughput Structural Studies

3.1. Method

A structural genomics target is a protein that is selected to determine its 3D structure. BSGC targets during the PSI-1 period include *Mycoplasma* proteins as well as their sequence homologues from other prokaryotes. In general, all rounds of target selection involved three common steps. Since almost all *MG* genes have their homologues in *MP*, each step was started with the set of 677 *MP* open reading frames (ORFs) described in the original annotation of the genome *(4)*. Each ORF was then augmented with a family of homologues from available, fully sequenced prokaryotic genomes

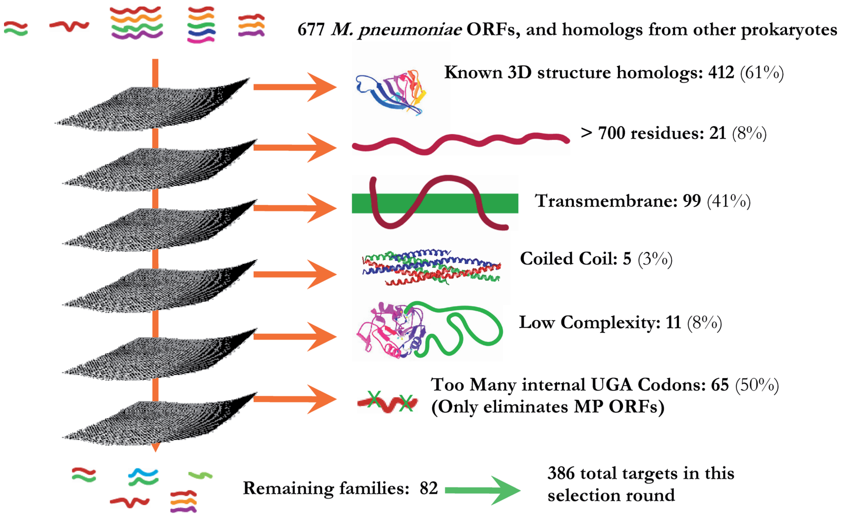
677 *M. pneumoniae* ORFs, and homologs from other prokaryotes
Known 3D structure homologs: 412 (61%)
> 700 residues: 21 (8%)
Transmembrane: 99 (41%)
Coiled Coil: 5 (3%)
Low Complexity: 11 (8%)
Too Many internal UGA Codons: 65 (50%)
(Only eliminates MP ORFs)
Remaining families: 82
386 total targets in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

Fig. 32.2 Target selection scheme for *Mycoplasma* genes used at BSGC. Criteria of "filtering" were changed among different rounds of target selection.

to make a target set. First, all target sets recognizably homologous to proteins of known structure in PDB 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 or low-complexity regions) were eliminated. Finally, specific targets were chosen from among proteins in the remaining target sets. The number of targets chosen per family varied among selection rounds. A flow diagram of target selection for a sample round is shown in Fig. 32.2.

3.2. Databases

The following databases were used in selection of targets:

1. MP: Each step was started with the set of 677 *MP* ORFs described in the original annotation of the genome *(4)*.
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 *(5)*, for which a structure has been solved by another structural genomics center. Sequences of BSGC targets that have progressed to the "Traceable Map" stage were also included, as this usually indicates the structure will soon be completed. The database was updated prior to each target selection round.
3. snr: the nonredundant set of protein sequences from Swiss-Prot *(6)*. All sequences were included in the swissprot, trembl, and trembl_new files (downloaded July 30, 2001 for round 2; November 30, 2001 for round 3; October 21, 2002 for rounds 4–5 and 7–8; and February 26, 2004 for round 6) from Swiss-Prot, which had been filtered with the SEG *(7)* and PFILT *(8)* programs using default options. The filtering was done to reduce the chance of profile corruption *(9)*, 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 web site (http://www.strgen.org).

3.3. Identification of Known Structures

At the beginning of each round of target selection, all *MP* proteins and their homologues 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 with the knownstr database using several sequence comparison tools such as PSI-BLAST *(10)*. PSI-BLAST position-specific scoring matrices (PSSMs) were constructed for each *MP* ORF using 10 rounds of searching the authors' "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 homologues, 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 to increase the possibility of identifying remote homologues, at the risk of a higher rate of corrupted PSSMs.
4. Because of the latter possibility, BLAST *(11)* and Pfam *(12)* in target selection rounds 3–6 were also used. 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 *(13)* 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. All ORFs that had a significant hit to a Pfam family that had also matched at least one known structure were eliminated from consideration.

3.4. Identifying MP Targets Predicted To Be Less Tractable for High Throughput Study

1. As the next step in each target selection round, *MP* proteins and domains that were likely to be predictably less tractable for high throughput study were eliminated. 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 May 24, 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 June 14, 1998), using the algorithm of Lupas *(14)*, for proteins with coiled coil regions, and two programs to iden-

tify transmembrane regions, TMHMM 2.0a *(15)* and PHDhtm *(16)* 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. Potential targets that were long and therefore likely to be challenging also were eliminated; 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, proteins annotated as ribosomal components were excluded, as these were expected to be unlikely to be stable in the absence of binding partners.

3.5. 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 described 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) 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.

3.6. 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 express truncated proteins in *E. coli*. In cases in which 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 multistep 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.

3.7. Target Deselection

The BSGC only seeks to solve structures for protein domains for which the structure cannot be reliably predicted via bioinformatic methods. Therefore, the authors deselect 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 *(17)*.

4. Protein Production

For this 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.

4.1. Cloning

For the past 2 years, BSGC has used an in-house version of the ligation independent-cloning (LIC) methodology *(18)*. 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) protease 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, there is a PCR-based method of preparing targets containing a C-terminal His_6 tag that can be used in cases in which 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, Fullerton, CA) robot: PCR reaction setup and cleanup, PCR product analysis by E-gel 96 (Invitrogen, Carlsbad, CA), LIC reaction and transformation, mini-expression setup, clone preservation in agar stab, and plasmid preparation.

4.2. Small-Scale Expression

After transformation into the expression host, two colonies are selected and grown in an autoinducing medium *(19)*. 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 *(20)*.

4.3. Large-Scale Preparation of Cell Paste

Previously Luria broth with isopropyl-β-D-1-thiogalactopyranoside (IPTG) induction was used. This required the manual addition of IPTG. For the past 2 years an auto-inducing medium (developed by William Studier from Brookhaven National Lab) has been used 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 autoinducing media are used: (1) ZYP: native medium, and (2) PASM: medium for labeling the target protein with seleno-methionine.

4.4. Protein Purification

Parallel purification is performed on three AKTA Explorer work stations (GE Healthcare, Piscataway, NJ). The authors have recently changed their protocol to the following: Five 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). This method takes 10 hours to complete.

4.5. Quality Control Assessments

All purified proteins undergo quality control steps 1–5 as listed in Table 32.1. One-dimensional NMR is performed on proteins smaller than 45 kDa that did not crystallize.

4.6. Protein Production Summary

As mentioned, BSGC is unique in that two minimal organisms (*MP* and *MG*) with the smallest genome size were chosen as the authors' target. The authors' targets have no sequence similarity to those of known structures. Even with a small starting pool of targets, a multipath approach eventually allowed the authors to produce most of their targets. For the minimal organism MP with 677 full-length predicted proteins, after filtering out the proteins that are structural homologues 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 homologues 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.

Table 32.1 Protein characterization

Parameters	Method
1. Purity	SDS/PAGE stained with Coomassie Brilliant Blue R
2. Monodispersity	Dynamic light scattering (DynaPro 99; Wyatt Technology, 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 A. Yakunin, University of Toronto, Toronto, Canada)
7. 1-D NMR	Bruker DRX 500 NMR spectrometer using an 11 (one-one) pulse sequence (D. Wemmer, University of California, Berkeley, CA)

5. Technical Development for Challenging Proteins

5.1. Heat Shock and High Salt Growth

Overexpression 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 *(21,22)*; 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 *(23,24)*. Simple heat shock before induction is known to enhance the solubility of some recombinant proteins produced in *E. coli (25)*. Some osmolytes behave as "chemical chaperones" by promoting the correct folding of unfolded proteins *in vitro* and in the cell *(26–29)*. These two elements, heat shock and high salt media, have been combined to increase the fraction of soluble protein produced from targets.

A protocol has been tested that combines heat shock and high salt growth *(30)*. 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 protein, 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 *(31)*.

5.2. On-Column Refolding

Inclusion body formation, as mentioned, 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; and (3) easily purified and solubilized. The main challenge is to convert inclusion bodies to properly folded, biologically active proteins. The authors have developed an on-column chemical refolding method *(32)* for insoluble His-tagged proteins expressed in *E. coli* partly based on the method described by Rozema and Gellman *(33)*. 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 *(34)*. 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, 10 of the PSI-1 targets from BSGC that expressed insoluble protein were purified using this method. Three of the 10 targets could not be refolded, but 30–100% refolding was obtained from the other seven. All refolded proteins were subjected to DLS analysis, and five of seven refolded proteins were monodisperse. Six of the seven refolded proteins were able to produce crystals of varying qualities.

5.3. Optimum Solubility 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. This step fails in about 25% of cases because the protein aggregates and precipitates; this adverse phenomenon is totally unpredictable. Inspired by a screen for NMR studies *(35)*, a screening method *(36)* was developed in which a panel of buffers as well as many additives were tested 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 was tested using the hanging-drop method and vapor diffusion equilibrium. After monitoring precipitation, the conditions leading to clear drops were selected for DLS characterization. For this part of the screen, only 24 µl of protein (with concentration ranging from 3 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 subsequently could be crystallized.

6. Structure Determination

The overall flow of the process for determining 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 multipath 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 multipath approach, the success rate was increased to approximately 27% (the corresponding success rate from clone to structure is about 5% and 16%, respectively) with additional multipath steps and automation. The overall pipeline at BSGC is schematically shown in Fig. 32.3.

6.1. 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 the sparse matrix crystallization screening method *(37)* and its commercially available variations (e.g., from Hampton Research, Aliso Viejo, CA). During this process, the Hydra Plus One

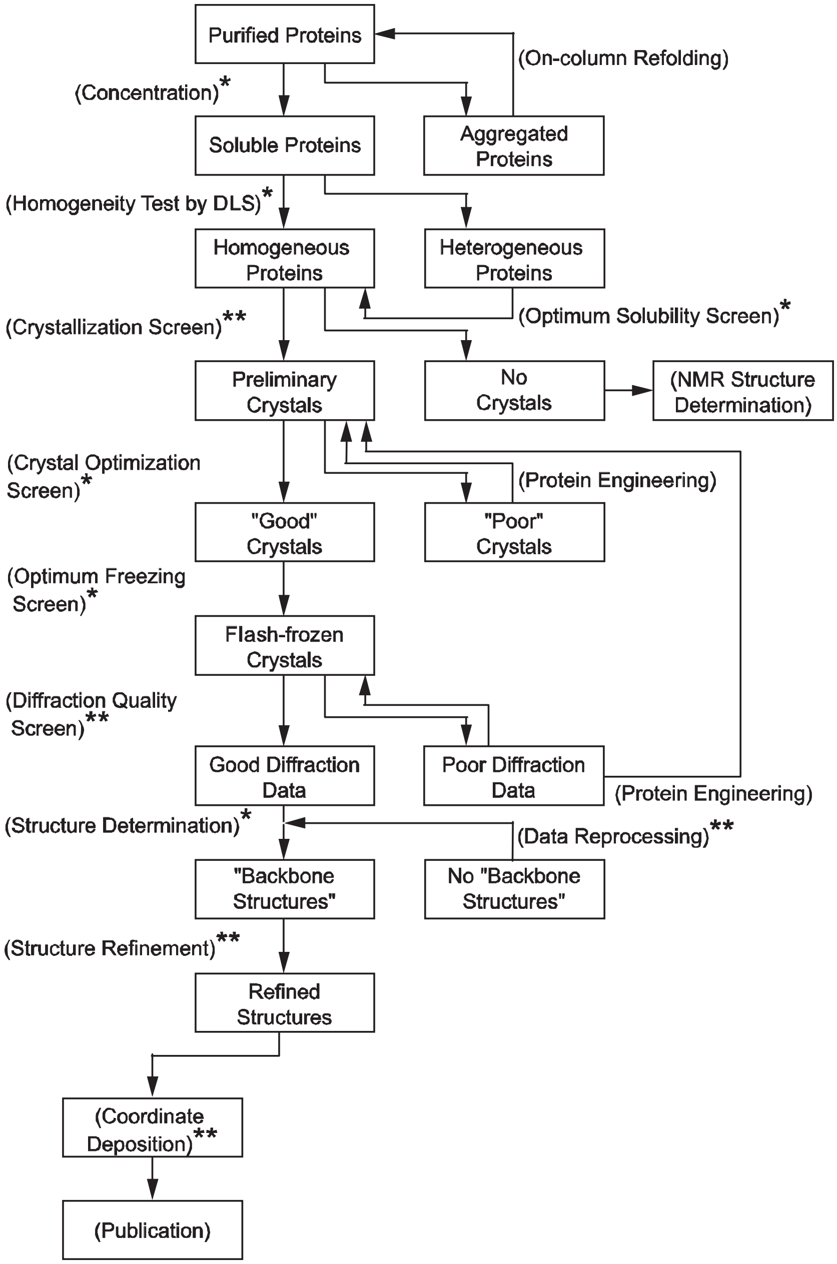

Fig. 32.3 Multipath flow diagram for the process from pure proteins to three-dimensional structures. The process for each step is in parenthesis. The automated steps are marked by **; steps that are partially automated or steps for which screens have been developed but have not been automated are marked by *.

(Matrix Technologies, Hudson, NH) and Phoenix Liquid Handling System (Art Robbins Instruments, Sunnyvale, CA) crystallization robots with 96-well plates are used. The authors routinely screen 4 × 96 crystallization conditions at two temperatures. Once one or more promising crystal hits are found, the hit conditions are optimized using a protocol developed by the authors to fine-tune the conditions to obtain good diffraction quality crystals. As a result of the on-column refolding step and optimum solubility screen, the success rate for the purified-protein-to-diffraction-quality-crystal process is about 27%.

6.2. Diffraction Data Collection

Many of the steps in diffraction data collection at Advanced Light Source, Lawrence Berkeley National Laboratory, 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.

6.3. Structure Solution

Once experimental data are collected, high throughput methods are applied to solve and complete a structure. The authors routinely use software developed for structural genomics efforts, such as HySS for substructure determination *(38)*. This software is part of the PHENIX package. The high level of automation that 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 *(39)*, MLPHARE *(40)*, or CNS *(41)* as dictated by data quality. In more challenging cases the SHARP *(42)* program is used. The pipeline is shown in Fig. 32.4.

The results of phasing are continued into phase improvement by density modification, using CNS *(41)*, DM *(43)*, and RESOLVE *(44)*. Visual inspection of the electron density map is used to determine whether more experimental data should be collected. For model building the authors use automatic software when possible. If data extend to 2.2 Å, the ARP/warp *(45)* 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 Å, a manual model building is used. Structure refinement and model completion makes use of the standard refinement tools: CNS and REFMAC *(46)*, automated water assignment, and manual rebuilding if necessary.

7. 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 homologues 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 homologues 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). The authors' overall success rates are shown in Table 32.2.

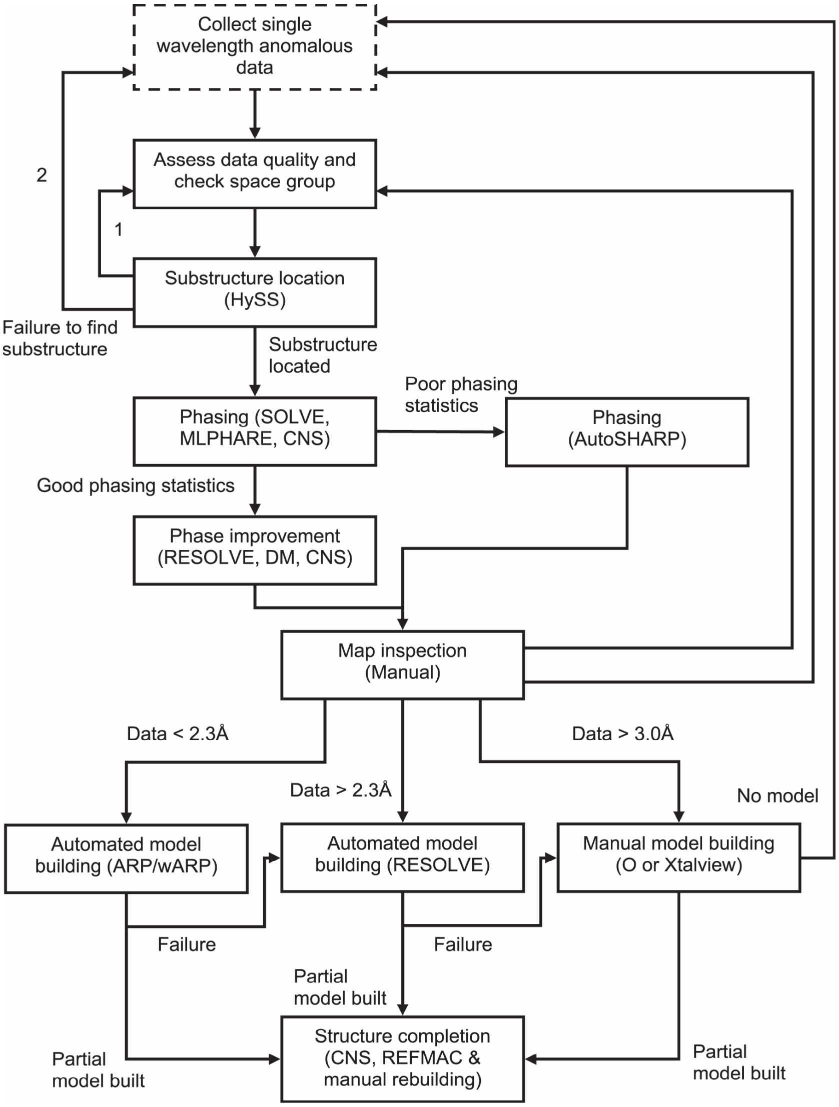
Collect single wavelength anomalous data
Assess data quality and check space group
2
1
Substructure location (HySS)
Failure to find substructure
Substructure located
Poor phasing statistics
Phasing (SOLVE, MLPHARE, CNS)
Phasing (AutoSHARP)
Good phasing statistics
Phase improvement (RESOLVE, DM, CNS)
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
Partial model built
Partial model built
Structure completion (CNS, REFMAC & manual rebuilding)

Fig. 32.4 Flowchart for structure solution, model building, and structure completion used by BSGC in the PSI pilot phase.

Table 32.2 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 homologues among proteins of known structures. Thus the authors 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 homologues"). The protein structures in categories 2 and 4 can infer possible molecular functions *(47)*.

8. The Protein Structure Space and the 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 *(48)*. Toward this objective, the authors have developed a computational process to represent all unique protein structures in a multidimensional space based on structural similarities and in 3D space for approximate visual representation of the multidimensional structural space.

8.1. 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 *(2,3)* that the protein structure space can be "mapped" in three dimensions as a visual approximation of multidimensional space of the protein structure space, 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 to reveal the demography of protein structures, and various structural information, functional information, and evolutionary information can be mapped and mined computationally, once such computational tools are developed.

8.2. Mapping of the Protein Structure Space

One of the major objectives of PSI is to obtain a broad coverage of the protein structure space (Fig. 32.5). To conceptualize the space, and derive new information from the demographic distribution of protein structures in the space, it is useful to define the space in terms of structural similarity. Calculating all pairwise structural similarity for all nonredundant protein structures (~2,000) in PDB, and converting them to structural dissimilarity scores, the authors were able to map the protein structure space in 3D space as a visual approximation of the full-dimensional representation (see Fig. 32.5). To accomplish this, the

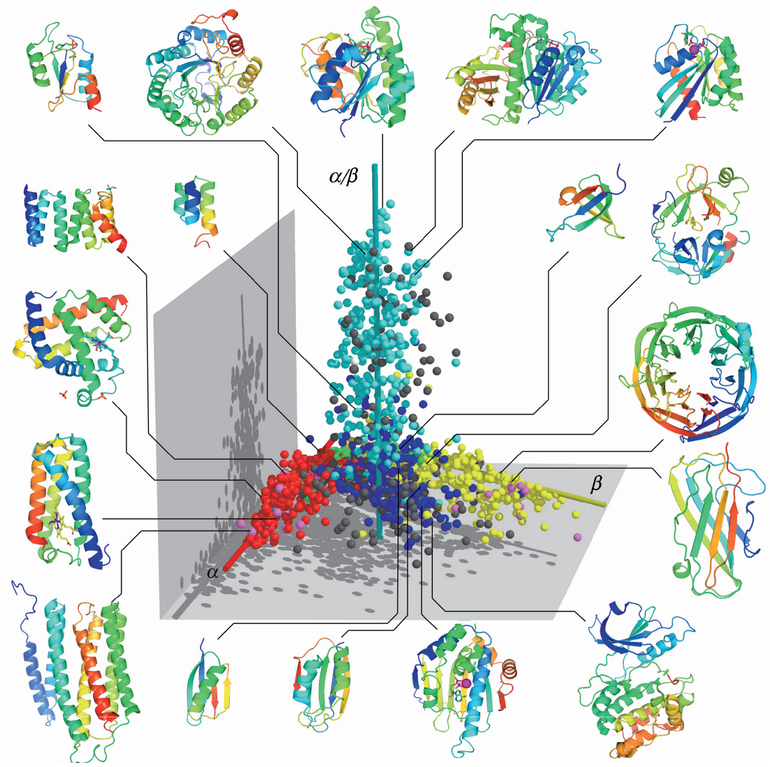

Fig. 32.5 Global representation of protein fold space. All together, 1,898 unique protein structure families are represented by spheres in the three-dimensional space. α, β, and α/β class structures follow three elongated directions denoted by α, β, and α/β feature axes *(3)*. Class designation for each structure family according to the SCOP *(49)* 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.

authors used the mathematical method known as multidimensional scaling *(2,3)*. 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 pairwise 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 supersecondary structure, elements generally increase along each feature axis, as indicated in Fig. 32.5. This suggests that these trends may be related to protein-fold evolution.

3. The three feature axes or the three "eigen vector" axes provide a completely general and objective way of classifying protein structures, thus providing a new demographic similar to library cataloguing. Furthermore, these structure features represented by axes 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 roughly within the "envelope" defined by the structural space originally found using approximately 2,000 nonredundant 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, to globally visualize the structural relationship among them, identify the regions of different structural population densities for suggesting additional new structures needed, and 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 fold/function relationship on a global level.

8.3. Structural Coverage of a Minimal Organism

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. At the end of PSI-1, the authors now have structural information for over 90% of the soluble proteins of this minimal organism (Fig. 32.6) *(50)*.

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 *(50)*.

8.4. Structural Families Found in the Minimal Organism

The unique structural families represented by all the soluble MG/MP proteins and their homologues are mapped on the protein structure space (Fig. 32.7).

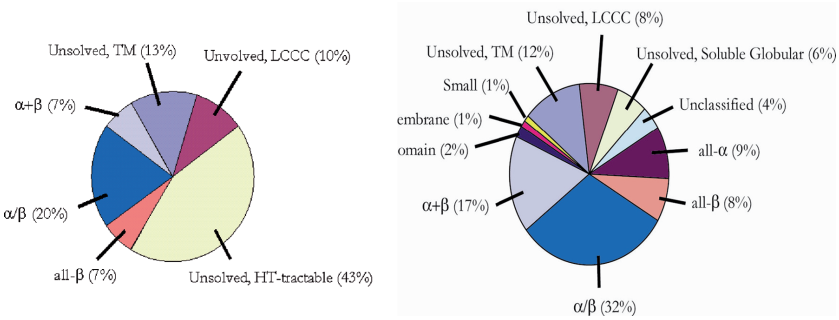

Fig. 32.6 Left. At the start of PSI-1, three-dimensional (3D) fold information was available for 34% of the proteins in *Mycoplasma genitalium*; the rest of the proteins belonged to membrane proteins (13%), low-complexity proteins (10%), and soluble proteins of unknown 3-D folds (43%). **Right.** By the end of PSI-1, 3D fold information was available for over 90% of soluble proteins in this minimal organism.

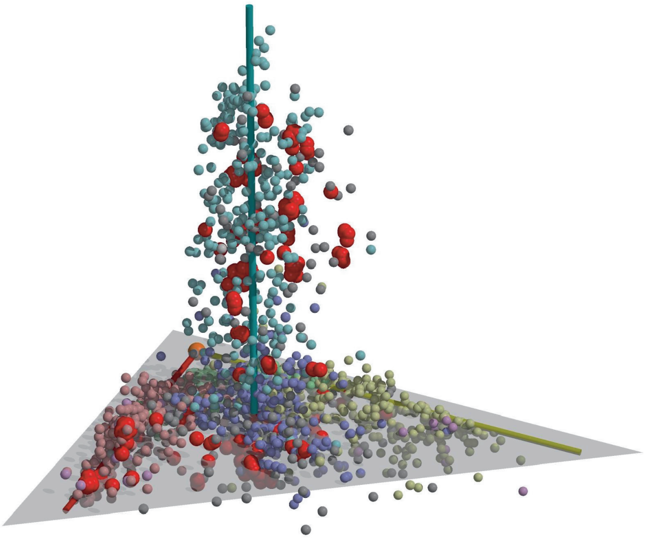

Fig. 32.7 Protein structure families determined at BSGC *(red)* mapped on the protein structure universe. Most of the BSGC structures had no sequence homologues in PDB structure database. About one half of them had new folds and occupy empty spaces in the protein structure universe, and the other half turned out to be “remote homologues” of structures of known folds and occupy the same or very close to preoccupied locations.

As expected, they are located within the envelope defined earlier by the 1,898 nonredundant structures of PDB. There appears to be a paucity in β-class proteins and more abundant usage of α/β-class proteins in these minimal organisms.

Acknowledgments

This work is supported by grants from the NIH (1-P50-GM62412 and 1-R01-GM073109). The authors are grateful to a large number of colleagues who participated in various aspects of BSGC’s PSI-1 program, such as high throughput cloning (H. Yokota and B. Gold) and expression (M. Henriquez and B. Martinez), large-scale production and characterization of proteins (C. Huang, Y. Lou, N. Oganesyan, and A. DeGiovanni), crystallization (J. Jancarik, I. Ankoudinova, and H. Hyun) and structure determination (D. Das, J. Liu, V. Oganesyan, and Q. Xian), and structural space mapping (S. Jun, G. Sims, J. Hou, and I.-G. Choi), with the guidance of S. Brenner, D. Wemmer, T. Earnest, D. McKay, and C. Hutchison, Jr.

References

1. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res*. **28,** 235–242.
2. Hou, J., Sims, G. E., Zhang, C., and Kim, S. -H. (2003) A global representation of the protein fold space. *Proc. Natl. Acad. Sci. USA* **100,** 2386–2390.

3. Hou, J., Jun, S.-R., Zhang, C., and Kim, S.-H. (2005). Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 3651–3656.
4. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C., and Herrmann, R. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae. Nucleic Acids Res*. **24,** 4420–4449.
5. Chen, L., Oughtred, R., Berman, H. M., and Westbrook, J. (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics* **20,** 2860–2862.
6. 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) The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. *Nucleic Acids Res*. **31,** 365–370.
7. Wootton, J. C. (1994) Nonglobular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem*. **18,** 269–285.
8. Jones, D. T., and Swindells, M. B. (2002) Getting the most from PSI–BLAST. *Trends Biochem. Sci*. **27,** 161–164.
9. Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., and Altschul, S. F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*. **29,** 2994–3005.
10. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search program. *Nucleic Acids Res*. **25,** 3389–3402.
11. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol*. **215,** 403–410.
12. 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) The Pfam protein families database. *Nucleic Acids Res*. **32,** D138–141.
13. Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics* **14,** 755–763.
14. Lupas, A. (1996) Prediction and analysis of coiled-coil structures. *Methods Enzymol*. **266,** 513–525.
15. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*. **305,** 567–580.
16. Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1995) Transmembrane helices predicted at 95% accuracy. *Protein Sci*. **4,** 521–533.
17. Chandonia, J. M., Kim, S. H., and Brenner, S. E. (2005) Target selection and deselection at the Berkeley Structural Genomics Center. *Proteins* **62,** 356–370.
18. Aslanidis, C., and De Jong, P. J. (1990). Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res*. **20,** 6069–6074.
19. Studier, W. (2005) Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif*. **41,** 207–234.
20. Nguyen, H., Martinez, B., Oganesyan, N., and Kim, R. (2004) An automated small-scale protein expression and purification screening provides beneficial information for protein production. *J. Struct. Funct. Genom*. **5,** 23–27.
21. Sachdev, D., and 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.
22. Harrison, S. C. (2004) Whither structural biology? *Nat. Struct. Mol. Biol*. **11,** 12–15.
23. Kempf, B., and Bremer, E. (1998) Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments. *Arch. Microbiol*. **170,** 319–330.
24. Bukau, B., and Horwich, A. L. (1998) The Hsp70 and Hsp60 chaperone machines. *Cell* **92,** 351–366.

25. Chen, J., Acton, T. B., Basu, S. K., Montelione, G. T., and Inouye, M. (2002) Enhancement of the solubility of proteins overexpressed in *Escherichia coli* by heat shock. *J. Mol. Microbiol. Biotech.* **4,** 519–524.
26. Samuel D., Kumar, T. K, Ganesh, G., Jayaraman, G., Yang, P. W., Chang, M. M., Trivedi, V. D., Wang, S. L., Hwang, K. C., and Chang, D. K., and Yu, C. (2000) Proline inhibits aggregation during protein refolding. *Protein Sci.* **9,** 344–352.
27. Yang, D. S., Yip, C. M., Huang, T. H, Chakrabartty, A., and Fraser, P. E. (1999) Manipulating the amyloid-beta aggregation pathway with chemical chaperones. *J. Biol. Chem.* **274,** 32970–32974.
28. Voziyan, P. A., and 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.
29. Diamant, S., Eliahu, N., Rosenthal, D., and Goloubinoff, P. (2001) Chemical chaperones regulate molecular chaperones *in vitro* and in cells under combined salt and heat stresses. *J. Biol. Chem.* **276,** 39586–39591.
30. Oganesyan, N., Ankoudinova, I., Kim, S.-H., and Kim, R. (2006) Effect of osmotic stress and heat shock in recombinant protein overexpression and crystallization. *Protein Express. Purif.* in press. [Au1]
31. 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.* **58,** 504–508.
32. Oganesyan, N., Kim, S.–H., and Kim, R. (2004) On-column chemical refolding of proteins. *PharmaGenomics* **4,** 22–26.
33. Rozema, D., and Gellman, S.H. (1996) Artificial chaperone-assisted refolding of denatured-renatured lysozyme: modulation of the competition between renaturation and aggregation. *Biochemistry* **35,** 15760–15771.
34. Daugherty, D. L., Rozema, D., Hanson, P. E., and Gellman, S. H. (1998) Artificial chaperone-assisted refolding of citrate synthase. *J. Biol. Chem.* **273,** 33961–33971.
35. Lepre, C. A., and Moore, J. M. (1998) Microdrop screening: A rapid method to optimize solvent conditions for NMR spectroscopy of proteins. *J. Biomol. NMR* **12,** 493–499.
36. 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.
37. Jancarik, J. and Kim, S. H. (1991) Sparse matrix sampling: a screening method for crystallization of proteins. *J. Appl. Cryst.* **2,** 409–411.
38. Grosse-Kunstleve, R. W., and Adams, P. D. (2003) Substructure search procedures for macromolecular structures. *Acta Cryst.* **D59,** 1966–1973.
39. Terwilliger, T. C., and Berendzen, J. (1999) Automated MAD and MIR structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **55,** 849–861.
40. Collaborative Computational Project, Number 4 (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **50,** 760–763.
41. 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., and Warren, G. L. (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Cryst.* **D54,** 905–921.
42. de La Fortelle, E., and Bricogne, G. (1997) Maximum-likelihood heavy-atom parameter refinement in the MIR and MAD methods. *Methods Enzymol.* **276,** 472–494.
43. Cowtan, K. (1999) Error estimation and bias correction in phase-improvement calculations. *Acta Cryst.* **D55,** 1555–1567.
44. Terwilliger, T. C. (2000) Maximum likelihood density modification. *Acta Cryst.* **D56,** 965–972.

45. Perrakis, A., Morris, R., and Lamzin, V. S. (1999) Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol*. **6,** 458–463.
46. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst*. **D53,** 240–255.
47. Kim, S. H., Shin, D. H., Choi, I. G., Schulze-Gahmen, U., Chen, S., and Kim, R. (2003) Structure-based functional inference in structural genomics. *J. Struct. Funct. Genom*. **4,** 129–135.
48. Kim, S.-H., Shin, D. H., Liu, J., Oganesyan, V., Chen, S., Xu, Q. S., Kim, J.-S., Das, D., Schulze-Gahmen, U., Holbrook, S. R., Holbrook, E. L., Martinez, B. A., Oganesyan, N., DeGiovanni, A., Lou, Y., Henriquez, M., Huang, C., Jancarik, J., Pufan, R., Choi, I.-C., Chandonia, J.-M., Hou, J., Gold, B., Yokota, H., Brenner, S. E., Adams, P. A., and Kim, R. (2005) Structural genomics of minimal organisms and protein fold space. *J. Struct. Funct. Genomics*. **6,** 63–70.
49. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. **247,** 536–540.
50. Chandonia, J. M., and Kim, S. H. (2006) Structural proteomics of minimal organisms: conservation of protein fold usage and evolutionary implications *BMC Struct. Biol*. **6,** 7–22.

Author Query:

[Au1]: update?