INVITED ARTICLE

Structural genomics: A pipeline for providing structures for the biologist

MARK R. CHANCE,[1,2,3] ANNE R. BRESNICK,[3] STEPHEN K. BURLEY,[4,5,10] JIAN-SHENG JIANG,[6] CHRISTOPHER D. LIMA,[7] ANDREJ SALI,[4] STEVEN C. ALMO,[1,3] JEFFREY B. BONANNO,[4] JOHN A. BUGLINO,[7] SIMON BOULTON,[8] HUA CHEN,[4,5] NARAYANAN ESWAR,[4] GUOSHUN HE,[4,5] RAYMOND HUANG,[1,2] VALENTIN ILYIN,[4] LINDA MCMAHAN,[4] URSULA PIEPER,[4] SOUMYA RAY,[4] MARC VIDAL,[8] AND LI KAI WANG[9]

[1]Center for Synchrotron Biosciences, Albert Einstein College of Medicine, Bronx, New York 10461, USA
[2]Department of Physiology and Biophysics, Albert Einstein College of Medicine, Bronx, New York 10461, USA
[3]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, New York 10461, USA
[4]The Rockefeller University, New York, New York 10021, USA
[5]Howard Hughes Medical Institute,
[6]Department of Biology, Brookhaven National Laboratory, Upton, New York 11973, USA
[7]Biochemistry Department and Structural Biology Program, Weill Medical College of Cornell University, New York, New York 10021, USA
[8]Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA
[9]Molecular Biology Program, Sloan-Kettering Institute, New York, New York 10021, USA

(RECEIVED November 13, 2001; FINAL REVISION January 8, 2002; ACCEPTED January 11, 2002)

Progress in understanding the organization and sequences of genes in model organisms and humans is rapidly accelerating. Although genome sequences from prokaryotes have been available for some time, only recently have the genome sequences of several eukaryotic organisms been reported, including *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, and humans (Green 2001). A logical continuation of this line of scientific inquiry is to understand the structure and function of all genes in simple and complex organisms, including the pathways leading to the organization and biochemical function of macromolecular assemblies, organelles, cells, organs, and whole life forms. Such investigations have been variously called integrative or systems biology and -omics or high-throughput biology (Ideker et al. 2001, Greenbaum et al. 2001, Vidal 2001). These studies have blossomed because of advances in technologies that allow highly parallel examination of multiple genes and gene products as well as a vision of biology that is not purely reductionist. Although a unified understanding of biological organisms is still far in the future, new high-throughput biological approaches are having a drastic impact on the scientific mainstream.

One offshoot of the high-throughput approach, which directly leverages the accumulating gene sequence information, involves mining the sequence data to detect important evolutionary relationships, to identify the basic set of genes necessary for independent life, and to reveal important metabolic processes in humans and clinically relevant pathogens. Programs such as MAGPIE (www.genomes.rockefeller.edu/magpie/magpie.html) compare organisms at a whole genome level (Gaasterland and Sensen 1996; Gaasterland and Ragan 1998) and ask what functions are conferred by the new genes that have evolved in higher organisms (Gaasterland and Oprea 2001). Concurrent with computational annotations of gene structure and function, thousands of full-length ORFs from yeast and higher eukaryotes have become available because of advances in cloning and other molecular biology techniques (Walhout et al. 2000a). Structural biologists have embraced high-throughput biology by developing and implementing technologies that will enable the structures of hundreds of pro-

Reprint requests to: Mark R. Chance, Department of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA; e-mail: mrc@aecom.yu.edu; fax: (718) 430-8587.
[10]Present address: Structural GenomiX, Inc., 10505 Roselle St., San Diego, CA 92121.
Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.4570102.

tein domains to be solved in a relatively short time. Although thousands of structures are deposited annually in the Protein Data Bank (PDB), most are identical or very similar in sequence to a structure previously existing in the data bank, representing structures of mutants or different ligand bound states (Brenner et al. 1997). Providing structural information for a broader range of sequences requires a focused effort on determining structure for sequences that are divergent from those already in the database. Although structure does not always elucidate function, in many instances (including the structures of two proteins reported here) the atomic structure readily provides insight into the function of a protein whose function was previously unknown. Typically, such functional annotations are based on homologies that are not recognizable at the sequence level but that are clearly revealed on inspection of the protein fold, identification of a conserved constellation of side-chain functionalities, or by the observation of cofactors associated with function (Burley et al. 1999; Shi et al. 2001; Bonanno et al. 2002).

Solving protein structures from diverse sequences: An international effort

When a protein with unknown structure has 30% or more sequence identity to a structure in the PDB, the structure of the unknown protein usually can be modeled accurately and useful functional information inferred (Marti-Renom et al. 2000). However, if the sequence identity is less than 30%, modeling the unknown sequence can lead to significant errors, even though the fold may be identified and functional information obtained. The ultimate goal of the National Institutes of Health (NIH)–funded structural genomics efforts (see below) is to determine the structure of at least one member of all sequence families sharing 30% or more sequence identity. The members of the family that are not determined directly will be modeled with significant accuracy (Sali 1998; Burley et al. 1999; Sali and Kuriyan 1999; Terwilliger 2000a; Sanchez et al. 2000; Shi et al. 2001). This process of using a template structure to calculate comparative models for related sequences is illustrated in Figure 1 (Marti-Renom et al. 2000). The template sequence of known structure (shown in green) is aligned with a homologous target sequence (red), and a model is then built for the target based on the template. The model is evaluated based on statistical criteria.

Structural genomics projects in the United States, Japan, and Europe have been established. Nine projects in the United States have received significant funding from the National Institute for General Medical Sciences. The nine scientific consortia (www.nigms.nih.gov/funding/psi.html) have pooled resources from several institutions (Table 1), and this group effort reflects a new paradigm for biology in which large consortia are formed to attack large problems, mimicking the high-energy physics and genome sequencing

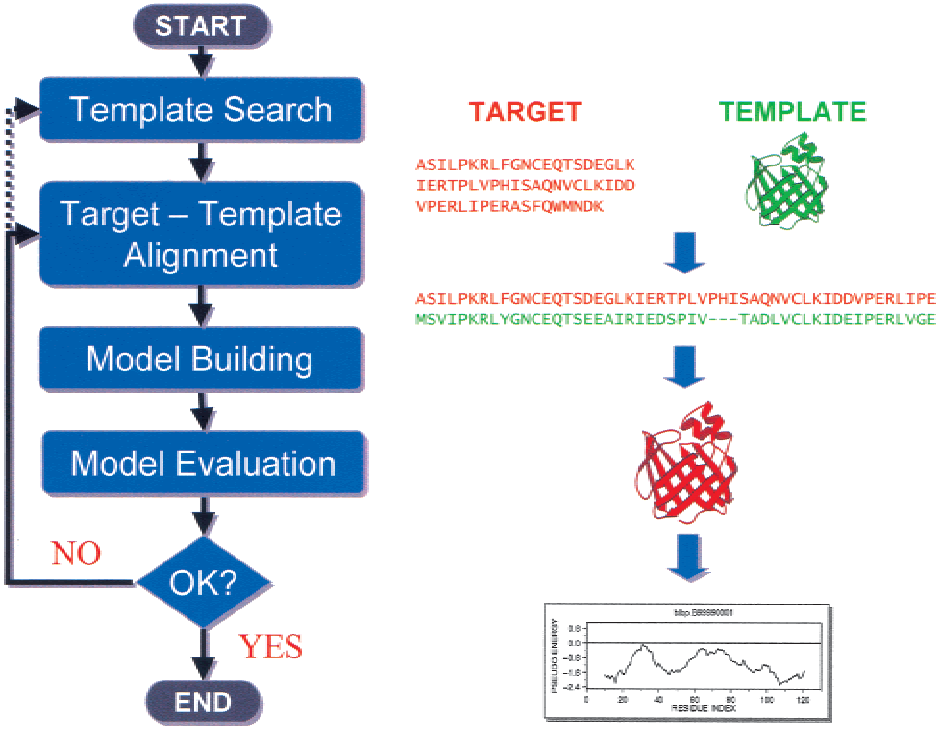

Fig. 1. Modeling target sequences based on a template structure (Marti-Renom et al. 2000).

Table 1. *NIH-funded structural genomics centers*

Center	Key organisms targeted	Involved institutions
Berkeley Structural Genomics Center	*Mycoplasma genitalium* and *Mycoplasma pneumoniae*	Lawrence Berkeley National Laboratory, University of California, Berkeley, Stanford University, University of North Carolina
Center for Eukaryotic Structural Genomics	*Arabidopsis thaliana*	University of Wisconsin and others
The Joint Center for Structural Genomics	*Caenorhabditis elegans*	The Scripps Institute, Stanford University, University of California San Diego
The Midwest Center for Structural Genomics	All three kingdoms of life	Argonne National Laboratory, Northwestern U. Washington University School of Medicine, University College London, University of Texas Southwestern Medical Center, University of Toronto, University of Virginia
New York Structural Genomics Research Consortium	Model organisms, human	Rockefeller University, Albert Einstein College of Medicine, Mt. Sinai School of Medicine, Weill-Cornell Medical, Brookhaven National Laboratory
Northeast Structural Genomics Consortium	Model organisms, human	Rutgers University, Columbia University, University of Toronto, Yale University
Structural Genomics of Pathogenic Protozoa Consortium	Pathogenic protozoa	University of Washington and others
TB Structural Genomics Consortium	*Mycobacterium tuberculosis*	Led by Los Alamos National Laboratory, more than 50 institutions involved
The Southeast Collaboratory for Structural Genomics	*Caenorhabditis elegans, Pyrococcus furiosus*	University of Georgia, Georgia State University, University of Alabama at Birmingham, University of Alabama at Huntsville, Oklahoma Medical Research Foundation, University of Oklahoma, Research Genetics, Massachusetts General Hospital, and Analiza, Inc.

projects. The New York Structural Genomics Research Consortium (NYSGRC; http://nysgrc.org), one of the nine NIH-funded centers, is a cooperative effort between the Albert Einstein College of Medicine, Rockefeller University, Brookhaven National Laboratories, Weill-Cornell Medical College, and Mount Sinai School of Medicine. The goal of the NYSGRC is to develop and implement high-throughput technology to identify, obtain, and model protein structures.

At this stage of the program, we are identifying bottlenecks in our structure discovery pipeline and developing or implementing technologies to remove these bottlenecks to increase the rate of structure determination. The pipeline includes: (1) target selection; (2) cloning the coding sequence of the targets into an appropriate expression vector; (3) sequencing the cloned gene to verify correct amplification of the coding sequence; (4) protein expression; (5) biophysical characterization of the expressed protein to confirm its identity and to establish the likelihood of crystallizability; (6) large-scale protein production for crystallization trials; (7) defining and refining crystallization conditions; (8) purification of selenomethione-labeled protein or identification of other suitable heavy-atom derivatives, and obtaining and freezing diffraction-quality crystals for X-ray crystallography; (9) collecting multiple wavelength anomalous dispersion (MAD) data at an X-ray synchrotron beam-line; (10) determining the phases of the reflections, building the model, and refining the structure; (11) making functional inferences from the structure, modeling unknown ORFs based on the identification of new sequence–structure relationships; and (12) disseminating our findings (Burley et al. 1999; Bonanno et al. 2002; Shi et al. 2002). Failures can occur at any step in the process for any individual target; thus, the pipeline is akin to a funnel, with a broad input and narrow output.

Dissemination to the scientific community occurs at three points in the pipeline. First, target lists are disseminated to minimize overlap between the centers and make available publicly the areas of focus. These target lists are available at http://targetdb.pdb.org/. Second, the results of each step in the pipeline for individual targets are provided through our publicly available Web-book (the integrated consortium experimental database, IceDB; http://nysgrc.org/). Third, structural coordinates are disseminated through the PDB. The Web has been extensively utilized to provide information on targets and structures; information from our consor-

tium is available in the public area of our Web site and for other consortia through www.structuralgenomics.org and through the Protein Data Bank Web site (www.rcsb.org/pdb/strucgen.html).

In this article, we recount recent progress of the NYSGRC in some of the pipeline steps identified above; particularly, we report on target identification and selection, development of a Web-based notebook for the consortium, and automating cloning, expression testing, and structure determination. We highlight this approach by describing two recently determined structures and by reporting the modeling of several thousand previously uncharacterized protein sequences based on the consortium's 27 new structures solved during its first year's effort.

Target selection for structural genomics

Structural genomics aims to structurally characterize most protein sequences by an efficient combination of experiment and modeling (Sali 1998, 2001; Burley et al. 1999; Sali and Kuriyan 1999; Marti-Renom et al. 2000; Sanchez et al. 2000; Vitkup et al. 2001). Central to the success of these efforts is effective target selection. There are a variety of target selection schemes, ranging from focusing on only novel folds to selecting all proteins in a model genome (Brenner 2000). Many of the target selection strategies of the NYSGRC are biologically based, providing a set of protein targets that are key actors in an interesting biological process. Such biologically interesting targets include all members of an enzymatic pathway, each protein in a large macromolecular complex, interacting proteins identified by two-hybrid screens or bioinformatics analysis, or a group of gene products that are up- and down-regulated in a biological process as determined by DNA-microarray techniques. For example, we detail below our examination of a set of cancer-related structural genomics targets that were identified by two-hybrid screening techniques as well as enzyme targets identified by bioinformatics analysis.

A model centric view of target selection, fulfilling the objectives outlined by the NIH, requires that targets be selected such that most of the remaining ORF sequences can be modeled with useful accuracy by comparative modeling. Even with structural genomics approaches, the structure of most proteins will be predicted by modeling and not determined directly by experiment. As discussed above, structural genomics needs to determine protein structures that have at least 30% identity to the sequences to be modeled (Marti-Renom et al. 2000; Vitkup et al. 2001). Recent estimates indicate that this cutoff requires a minimum of 16,000 targets to cover 90% of all protein domain families, including those of membrane proteins (Vitkup et al. 2001). These 16,000 structures subsequently will allow us to model many more proteins. In practice, for the new structures so far determined by the NYSGRC, an average of 100 protein sequences without any prior structural characterization could be modeled at least at the fold level (http://nysgrc.org/). The ability to leverage structure determination by protein structure modeling illustrates and justifies the premise of structural genomics.

Target selection for structural genomics of enzymes

Our consortium has multiple target selection strategies; one of which involves targeting enzymes. Enzymes are typically soluble proteins of metabolic and biomedical interest, either as drug targets in pathogenic organisms or in understanding metabolic function in human or animal models. We have prepared a conservative list of protein families that contain human enzymes of unknown structure. First, all sequences with an annotated enzyme classification (EC) number were extracted from the TrEMBL database (9/1/01), resulting in 19,382 presumptive enzymes from a wide variety of organisms. This list included human enzymes in 204 classes with unique EC numbers. For each of the 204 representative human enzymes, homologs from 10 other organisms with greater than 30% sequence identity were selected. The choices reflected the current or expected availability of full-length cDNAs for *Bacillus subtilis*, *C. elegans*, *D. melanogaster*, *Escherichia coli*, *Methanococcus jannaschii*, *Mus musculus*, *S. cerevisiae*, *Streptococcus aureus*, *Streptococcus pneumoniae*, and *Thermotoga maritima*. The homologs were identified from the PSI-BLAST profiles in ModBase (Pieper et al. 2002). The resulting 204 families contain 903 enzymes; all of which were deposited into the IceDB target tracking system, which is our on-line database and laboratory notebook (see below) and annotated. The final target list was further refined by examining the following criteria. (1) Domain structure was examined by conducting sequence alignments of the families. This inspection identified sequences that actually were not closely related but had the same enzyme annotation and managed to pass through the loose PSI-BLAST screening at the enzyme family level. (2) Length of the target sequence was considered. Enzyme targets of 500 amino acids in length or less are preferred. (3) The presence of predicted *trans*-membrane spanning regions also was examined and was avoided. In addition, all enzymes of known structure and enzymes that can be related to a protein of known structure with a PSI-BLAST *E*-value greater than 10^{-4} over any segment in their sequences were not selected as targets. The current list of selected targets contains ~300 enzymes from ~100 EC classes, with each class containing a single human sequence and multiple homologs from the selected organisms. These targets will be subjected to automated cloning and publicized on our Web site.

IceDB: The consortium Web book

IceDB is a multifunctional Web resource of the NYSGRC for depositing, filtering, tracking, and publicizing of struc-

tural genomics targets (http://nysgrc.org/). First, IceDB archives the preliminary target sets and facilitates manual filtering to obtain final target sets for X-ray crystallography. Second, IceDB is also a Laboratory Information Management System for entering, storing, querying, and displaying the progress made with each of the selected target proteins. The software has been designed with flexibility in mind, and it is relatively easy to add functionalities as needed. Finally, IceDB is the public face of the NYSGRC as it reports the progress of the NYSGRC and maintains an up-to-date list of the selected targets in the XML format for the consolidated structural genomics Web site at the Protein Data Bank. IceDB is closely linked to our other resources used in target selection and structure-based annotation, including ModBase.

ModBase: A comprehensive database of comparative protein structure models

ModBase (http://guitar.rockefeller.edu/modbase) is a relational database of annotated comparative protein structure models for all available protein sequences matched to at least one known protein structure (Sanchez et al. 2000; Pieper et al. 2002). It is a critical tool for target selection by the NYSGRC and is used to calculate comparative models of newly solved structures. Comparative models in ModBase are calculated using ModPipe, an entirely automated software pipeline for large-scale comparative protein structure modeling (Sanchez and Sali 1998). ModPipe relies on PSI-BLAST (Altschul et al. 1997) and IMPALA (Schaffer et al. 1999) for fold assignment, and the MODELLER package for sequence–structure alignment, model building, and model assessment (Sali and Blundell 1993). Fold assignments and models for a fraction of these sequences are considered unreliable. The folds of the models are assessed by computing an energy-based model score that uses a statistical energy function, sequence similarity with the modeling template, and a measure of structural compactness (Sanchez and Sali 1998). Tests with known structures have shown that models with scores from 0.7 to 1.0 have the correct fold at a 95% confidence level. ModBase uses the MySQL relational database management system for flexible and efficient querying and the ModView Netscape plugin for viewing and manipulating multiple sequences and structures. It is updated regularly to reflect the growth of the protein sequence and structure databases and to add improvements in the software for calculating models. For ease of access, ModBase is organized into different data sets. The largest data set contains models for domains in 304,517 of 539,171 unique protein sequences in the complete TrEMBL database (3/23/01); only models based on significant alignments (e.g, PSI-BLAST *E*-values greater than 10^{-4}) and models assessed to have the correct fold are included.

Many ModBase data sets are created by the Web server for automated comparative protein structure modeling, ModWeb (http://guitar.rockefeller.edu/modweb). ModWeb provides a Web interface to ModPipe and takes as input either a set of sequences or a protein structure. For all input sequences, models are calculated when a potentially related known protein structure is found in the PDB. For an input protein structure, models are produced for all the detectably related protein sequences in a comprehensive nonredundant sequence database (see above). ModBase provides convenient storage and access to the models calculated by ModWeb.

Production and testing of expression clones

A key feature of the structural genomics pipeline is the automated production of expression vectors from identified targets. Within the NYSGRC, we have established a centralized cloning facility that is responsible for operating an automated platform for all of the molecular biological steps required to subclone ORFs from genomic DNAs and/or cDNA libraries, insert these coding sequences into expression vectors, transform *E. coli*, and test the resulting expression strains for production of soluble protein. Our initial approach is based on the use of topoisomerase-mediated, directional flap ligation of a blunt-ended PCR product into the Invitrogen pET100/D-TOPO Vector, which creates a fusion protein bearing an N-terminal His_6-tag followed by a polio viral protease cleavage site followed by the protein of interest. We have implemented an additional vector for recombinant protein expression based on N-terminal fusions with a yeast form of SUMO, a small ubiquitin-like modifier that appears to confer solubility to the recombinant fusion protein (Mossessova and Lima 2000). The pSUMO system utilizes an N-terminal His_6-tag SUMO fusion with the respective target sequence. The protein is expressed, purified by metal affinity chromatography, and liberated from the His_6-SUMO fusion by cleavage with a modified version of the desumoylating enzyme Ulp1.

A Beckman Biomek FX Robotic Platform has been programmed to perform all of the steps required to go from PCR primers to transformed BL21(DE3) Star in 96-well format with bar code tracking of sample and reagent plates. Some steps are conducted off-line with multichannel pipetting. Small-scale (1 μg) purification of recombinant proteins then is performed with Millipore Metal Chelating ZipTips, loaded with Ni^{2+} ions. The resulting purified recombinant proteins can be spotted onto a matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) sample plate.

Our initial experiences with this technology platform are extremely encouraging. The first cloning candidate (NYSGRC ID T136) subjected to the entire process yielded purified recombinant protein with a measured molecular mass

within 12 mass units of the predicted mass of 22,845.8 (observed mass = 22,834.4). After cleavage with polio viral protease, we obtained a measured molecular mass within 73 mass units of the predicted mass of 17,990.5 (observed mass = 17,917.6). Thus, we can rapidly confirm soluble expression of the desired protein with the correct molecular mass and removal of the His_6 purification tag without DNA sequencing of the expression plasmid. Expression strains meeting these criteria are transferred to one of the five decentralized protein production/crystallization teams located at each of the five participating institutions. However, proteins that crystallize and proceed to structure determination will need to be checked for mutations that do not significantly affect the molecular mass, such as Ile/Leu, Gln/Glu, or Asp/Asn. For this restricted number of cases, we will perform DNA sequencing with plasmid DNA that has been stored with appropriate bar code registration.

At present, one molecular biologist working with the robotics technician can work through the entire process of cloning/transformation/insert testing/retransformation/expression testing for 96 clones in less than 2 weeks. Preliminary results showed a 90% success rate in producing BL21(DE3) Star cells transformed with the desired expression vector using the above enzyme targets as input to the cloning strategy. Solubility tests demonstrated that 75% of these robotically engineered expression clones yielded soluble proteins with correct apparent masses (as judged by gel electrophoresis). In the long term, small-scale, automated biophysical characterization will include domain mapping by limited proteolysis and MALDI-MS and detection of nonspecific aggregates using dynamic light scattering. The results of these studies will be used to guide target redesign within the NYSGRC.

High-throughput expression and solubility testing of *C. elegans* gateway ORF clones

The availability of cloned genes representing full-length ORFs from higher organisms provides a unique opportunity for including eukaryotic targets in the structural genomics pipeline. In addition, targeting genes involved in specific biological processes provides a biological focus that can nucleate collaborations among different laboratories. Vidal and coworkers have developed high-throughput methods for cloning thousands of eukaryotic genes using the GATEWAY system (Walhout et al. 2000a) and have used cloned genes from *C. elegans* to perform two-hybrid analyses (Walhout and Vidal 2001) to screen for protein–protein interactions in several interesting biological processes including vulval development (Walhout et al. 2000b), the 26S proteosome (Davy et al. 2001), and cell cycle control and DNA damage (Boulton et al. 2002). For the latter screen, more than 70 *C. elegans* genes implicated in cell cycle control and DNA damage based on homology with yeast and human genes known to be involved in these processes were used as the baits. More than 90 interactors were identified, and both the bait and interactor genes were subcloned using the in vitro phage λ-based recombination technology into the appropriate GATEWAY destination vectors (Life Technologies Inc.) for the production of both His_6 and maltose binding protein (MBP) –tagged fusion proteins. One hundred sixty-seven full-length His_6-tagged and 86 (otherwise identical and full-length) MBP-tagged proteins were assayed for expression and solubility in *E. coli* using a 96-well plate format. Working by hand, we processed 96 constructs in 2–3 days not including gel analysis. Of the 253 constructs screened, 143, which had molecular masses ranging from 10 to 140 kD (not including the affinity tag), displayed moderate to high expression levels (Table 2). Using a simple lysis procedure (see Materials and Methods), we found that 31 of the 143 expressed fusion proteins exhibited significant solubility. Our analysis showed that 61% of the expressed MBP fusions were soluble, whereas only 9% of the expressed His_6 fusions were soluble, consistent with previous studies showing that MBP increases the stability and solubility of proteins expressed in *E. coli* (Kapust and Waugh 1999). More importantly, this analysis indicates that at a minimum 9% of the *C. elegans* genes examined are soluble in an *E. coli* expression system without any special efforts at solubilization. Of significant interest to the NYSGRC are the 112 fusion proteins that exhibit high levels of expression but that are insoluble under the current lysis conditions. These proteins are excellent candidates for the high-throughput solubilization/refolding techniques that currently are being developed by the NYSGRC. Also, Table 2 lists novel targets; these represent genes for which no PDB entry exists that is greater than 30% identical to the target in question.

The high-throughput screen of the *C. elegans* constructs, revealed both the strengths and weaknesses of the GATEWAY cloning system for structural genomics. The in vitro recombination reaction permitted the rapid transfer of the *C. elegans* ORFs into multiple destination vectors, which maximized the number and types of expression constructs that could be examined for expression and solubility. In fact, it was entirely unexpected that such a significant fraction of the *C. elegans* genes would produce robust amounts of protein under the simple bacterial growth and expression

Table 2. *Summary of C. elegans high-throughput screen*

	Total	MBP	His_6
Clones screened	253	86	167
Expressing clones	143	34	109
Soluble clones	31	21	10
Novel targets	11	4	7

(MBP) maltose binding protein.

protocols utilized, and this bodes well for attempts to express many eukaryotic genes in bacterial systems. However, we encountered three basic problems with the GATEWAY system (as it is commonly used in high-throughput proteomics efforts) with respect to protein crystallization. The stop codons typically are provided by the destination vector, and the precise position varies depending on the particular reading frame. This results in the insertion of at least eight additional amino acids at the C terminus caused by the recombination site and vector sequences. Second, eight additional amino acids are inserted between the N terminus of the target and the affinity tag caused by the recombination site. These extra amino acids are likely to be a significant problem because additional unstructured regions at the N or C terminus may inhibit crystallization (Shi et al. 2001).

Last, for the N-terminal fusions there is either no protease cleavage site (His_6-tagged proteins), or the protease cleavage site lacks sufficient specificity (MBP-tagged proteins), thus precluding reliable release of the intact protein after removal of the affinity tag. On the basis of these observations, we anticipate that the GATEWAY system as presently configured for these studies will be of value for high-throughput expression and solubility testing; however, targets identified in this screen will have to be recloned if they look promising. However, the entry vector strategy could easily be modified to be optimal for subsequent large-scale protein purification and crystallization efforts of structural genomics projects.

A new bacterial non-haem Fe(II)-dependent oxygenase family

We report the structure of two proteins recently solved by the consortium. The first is the *E. coli* Gab protein (NYSGRC ID T130) that belongs to a small but highly conserved sequence family annotated as an ORF in the γ-amino-butyric acid (GABA) operon of *E. coli* and *Salmonella typhimurium*. The function of the Gab protein is not known, and its sequence relationship to other proteins has not revealed functional insight. This investigation is an excellent example of the value of structure in providing insight into possible function. Our analysis of the *E. coli* Gab protein reveals it is a member of the non-haem iron (II)-dependent oxygenase superfamily, a family of enzymes that include clavaminic acid synthase (CAS; Zhang et al. 2000), deacetoxycephalosporin C synthase (DOACS; Valegard et al. 1998), and isopenicillin N synthase (IPNS; Roach et al. 1997), enzymes that catalyze β-lactam ring formation in antibiotics and known inhibitors of β-lactamases (Baldwin and Abraham 1988).

T130 is a member of one family of non-haem Fe(II)-dependent oxidases/oxygenases and utilizes an octahedral Fe(II) center that participates in the oxygenase chemistry. The catalytic iron is coordinated by side chains from three amino acid residues, two histidine residues, and a carboxylate donated by either aspartic or glutamic acid. Two additional coordination sites for iron can be provided by 2-oxoglutarate (2-OG) as in CAS or through direct coordination of a substrate peptide as is observed in IPNS. This leaves open a coordination site for dioxygen binding, an essential reactant in oxygenase reactions. The X-ray structure analysis of *E. coli* Gab reveals that it is a member of the 2-OG Fe(II)-dependent family of oxygenases and likely represents a unique and small family of oxygenases found in the enteric bacteria *E. coli* and *S. typhimurium*.

Detailed structure of *E. coli* Gab

E. coli Gab was expressed, purified, crystallized, and characterized by X-ray crystallography (see Materials and Methods). The structure of the Gab protomer is composed of a mixed seven-stranded β sheet (β1, β2, β3, β14, β7, β16, and β5). Five of these strands (β3, β14, β7, β16, and β5) and three additional strands (β8, β15, and β6) are involved in formation of a distorted jelly roll (Fig. 2A,B). The β sheet and jelly roll appear stabilized in part by helices α4 and α5. The *E. coli* Gab protein behaves as a tetramer in solution as analyzed by gel filtration chromatography. Crystallographic analysis of *E. coli* Gab in two spacegroups (I422 and $P42_12$) reveals tetrameric organization between Gab protomers, suggesting the crystallographic tetramer is similar to that observed in solution (Fig. 2A,B). The tetramer is formed through interactions between helices α1–3 from one protomer and α6–7 from an adjacent protomer. The observed interactions are mixed and involve hydrophobic, hydrogen-bonded, and salt-bridging interactions. Some of the many interactions include Tyr 149 and Leu 150 interaction with Tyr 59′, Tyr 243 and Phe 253 interaction with Phe 65′, Asp 264 interaction with Arg 66′, and Glu 158 interaction with Lys 60′. The calculated total buried accessible surface area for each protomer–protomer interaction is 2410 $Å^2$ (Nicholls et al. 1991). The tetrameric structure of *E. coli* Gab buries ~8620 $Å^2$ in total. The Gab oligomer appears to be unique to this enzyme subfamily because other structurally characterized members of the non-haem Fe(II)-dependent oxygenase family are monomeric, and the conserved elements involved in protomer interactions are limited to *E. coli* and *S. typhimurium* family members.

The fold of the Gab protomer places it within the clavaminate synthase-like family (SCOP) (Murzin et al. 1995). The three structurally characterized members within this family include clavaminate synthase (PDB code 1DRT; Zhang et al. 2000), isopenicillin N synthase (IPNS, PDB code 1BK0; Roach et al. 1997), and deacetoxycephalosporin C synthase (DOACS, PDB code 1DCS; Valegard et al. 1998). The structure of Gab can be aligned over 144 amino acids with other Fe(II)-dependent oxygenases to an RMSD of 3.0 Å over 265 amino acids (Cα-atoms) for CAS, 4.2 Å

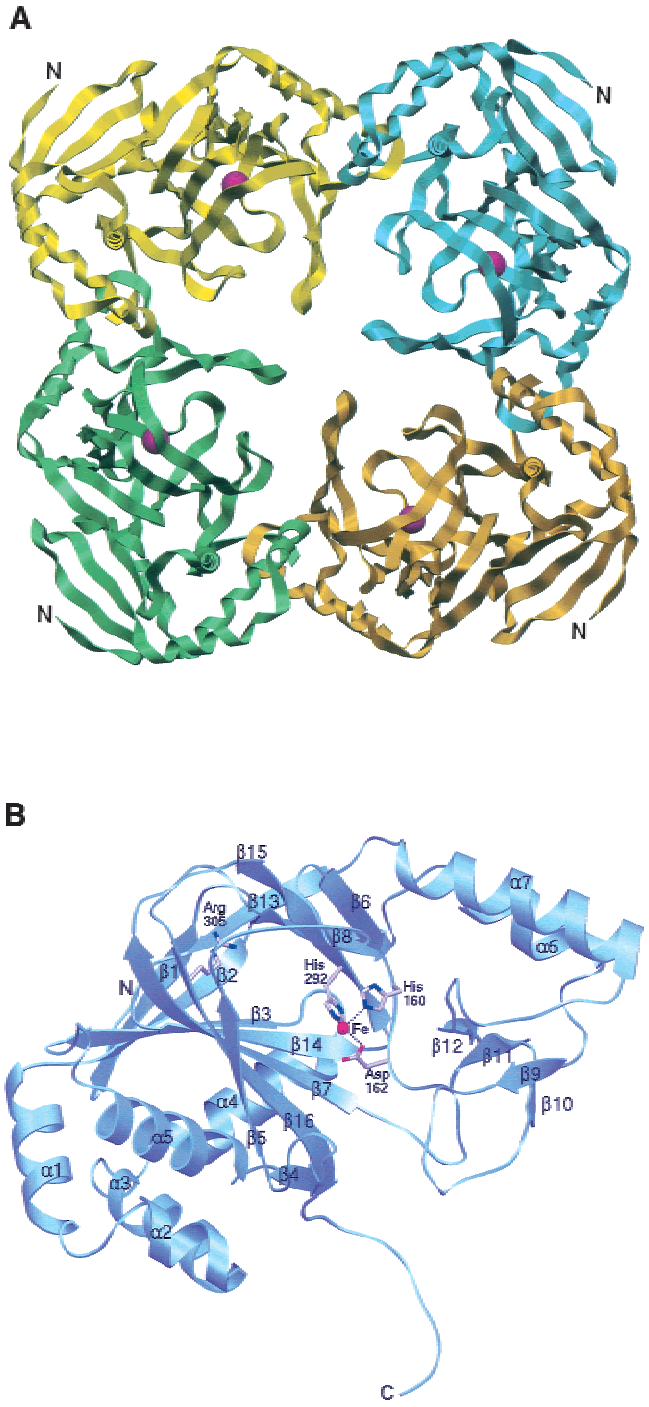

Fig. 2. Ribbon diagram of *E. coli* Gab. (*A*) Oligomeric architecture of the Gab tetramer. Monomers are colored yellow, blue, gold, and green. The N terminus is labeled N for each protomer. Large magenta spheres mark the Fe(II) ion in each protomer. Each active site is well separated and distinct. Oligomeric interactions are mediated primarily between α1–3 and α5–6 between respective protomers. (*B*) Protomer of the Gab protein. N- and C-terminal residues are denoted with an N and C, respectively. Secondary structural elements are denoted α1–7 and β1–16 corresponding to the sequence alignment presented in Fig. 3. Active site Fe(II) and residues His 160, Asp 162, His 292, and Arg 305 are labeled. Graphics prepared using SETOR (Evans 1993).

for DOACS, and 3.9 Å for IPNS as determined by a structural homology search with the program DALI (Holm and Sander 1993). The general features conserved within this group of enzymes are a distorted jelly roll and active site residues that coordinate Fe(II) for catalysis. Although CAS, IPNS, and DOACS share striking structural and chemical similarities with *E. coli* Gab, the sequence identity between family members and the divergent oligomeric structure of Gab suggest that Gab is not a functional homolog for IPNS, CAS, or DOACS. This is illustrated by a structure-based sequence alignment between the four structural representatives within this enzyme family, which shows that there is little sequence identity and several large gaps between family member sequences (Fig. 3). Because of the unique character of the Gab sequence family, structure-based sequence analysis software such as PFAM and Modeler do not reveal relationships to CAS, IPNS, DOACS, or any other protein families (data not shown).

E. coli Gab active site

As observed in other non-haem Fe(II)-dependent oxygenases, the Gab protein coordinates an iron ion with two histidine residues (His 160 and His 292) and a side-chain carboxylate (Asp 162). The metal ion appears coordinated in an octahedral geometry with water molecules occupying the remaining positions. In CAS, the closest structural homolog of Gab, two coordination sites are occupied by 2-OG (Zhang et al. 2000). We have grown crystals of Gab under anaerobic conditions (argon atmosphere) in the presence of 20 mM 2-OG); however, diffraction analysis does not identify bound 2-OG in the Gab active site. Density that was not attributed to side-chain or main-chain atoms existed in proximity to the iron center in the same approximate position for peptide ligands observed in structures for CAS, IPNS, and DOACS, suggesting that copurification of a putative ligand for this enzyme may have occurred. This material remains unidentified. Interestingly, a conserved arginine (Arg 305) within the superfamily also is conserved in Gab. This residue is utilized in binding substrate carboxylate groups in IPNS, DOACS, and CAS.

The structural approaches described here have identified a unique class of non-haem iron-dependent oxygenases in *E. coli* and *S. typhimurium*, suggesting these organisms may be capable of synthesizing compounds similar to those generated by CAS, IPNS, and DOACS. The function for Gab in *E. coli* is not known. Its genetic linkage to enzymes involved in γ-amino-butyric acid (GABA) metabolism (GabD or succinic semialdehyde dehydrogenase) and its structural similarity to the CAS 2-OG-dependent oxygenase is suggestive because 2-OG plays a central role in GABA and L-glutamate metabolism. The structural analysis presented here has revealed chemical properties of the active site that will aid in elucidating the biochemical function for *E. coli*

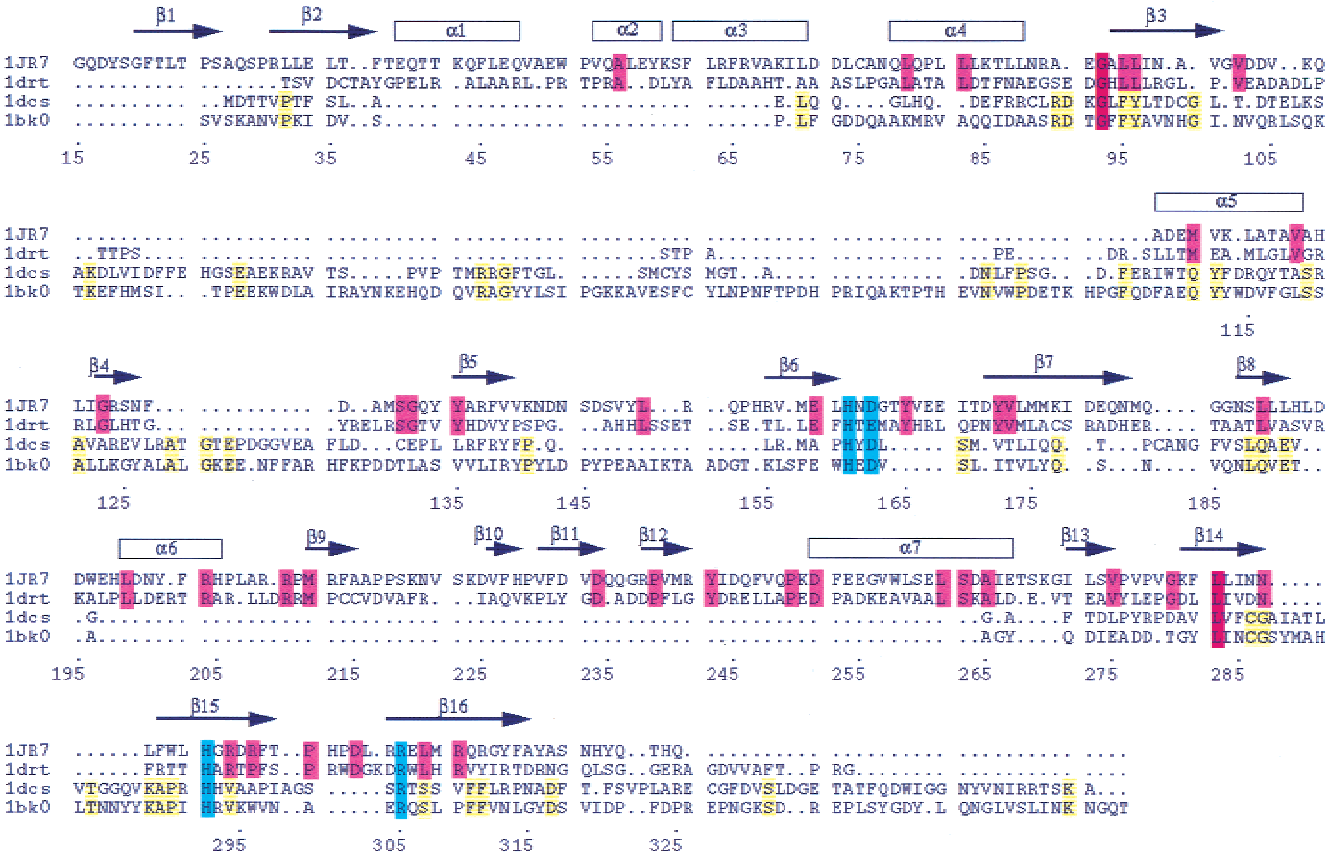

Fig. 3. Structure-based sequence alignment for Gab with other non-haem iron oxygenases. Structure-based sequence alignment produced by PrISM (Yang and Honig 1999). 1JR7, 1drt, 1dcs, and 1bk0 represent PDB codes for *E. coli* Gab, clavaminate synthase 1 (CAS), deacetoxycephalosporin C synthase (DOACS), and isopenicillin N synthase (IPNS), respectively. Highlighted residues in alignment are color-coded blue (active site residues) and magenta if conserved across all family members, yellow if conserved between DOACS and IPNS, and orange if conserved between CAS and Gab. Sequence identities in the structure-based alignment revealed 15.6% identity between Gab and CAS, 10.3% identity between Gab and DOACS, 5.0% identity between Gab and IPNS; 18.3% sequence identity was found between DOACS and IPNS, 7.4% identity between DOACS and CAS, and 9.8% identity between CAS and IPNS. Secondary structural elements (above) and sequence numbering (below) are shown for *E. coli* Gab in the alignment.

Gab. Thus, although the function of Gab was not known, its inclusion in the structural genomics pipeline and the solved structure provide important clues to understanding its function and point the way toward future research.

Automated structure determination platform

An automated structure determination platform (http://asdp.bnl.gov) has been established by the NYSGRC for high-throughput structure determination (Fig. 4). ASDP provides an integrated interface to the most frequently used crystallographic programs, databases, and Web interfaces. Various publicly accessible software packages are organized as a production pipeline that provides a highly efficient computational environment spanning all steps of structure determination, including data collection; initial phasing, modeling, refinement, and PDB submission (http://asdp.bnl.gov/asda/ASDP/index.html). This platform provides Web-based interfaces and format-conversion tools to allow an easy flow of information among programs. The platform is implemented on a substantial and expandable computing server (http://asdp.bnl.gov/asda/About/asdp_server.html). The computing resources are accessed through a Web-based graphical user interface and an easy-to-use Unix-like file system. The registered users are provided a Web-based home data directory with disk space. X-ray data collected at synchrotron beam lines can be linked directly to a users Web-home within ASDP. The user also can upload and download files between their own workstation and the Web-home. Molecular and structural information is imported directly from an internal database. Crystallographic information (i.e., space-group, unit cell, etc.) determined during data collection is extracted directly from the data files. After setting up necessary experimental data files, ASDP automatically per-

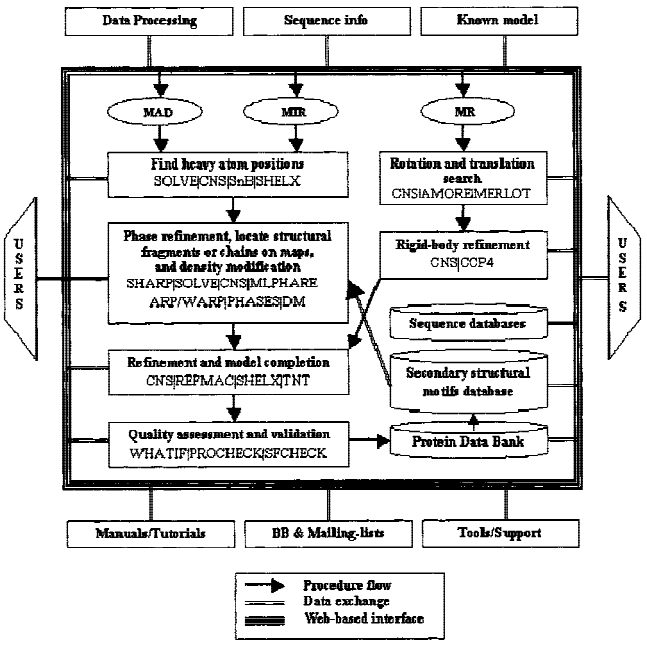

Fig. 4. The flowchart of Automated Structure Determination Platform.

forms appropriate data conversions required to use the various crystallographic programs and writes script files in the user's directory for running programs. Users are presented with a list of steps/programs, each with suggested input/script files for submitting jobs. The jobs run on the large computing cluster (Linux or SGI) available to the ASDP user community and the progress of each structure determination is archived in an internal database. The mmCIF format is supported for internal and external data exchange.

Structure determination of Target P097 using ASDP

ASDP was tested in the structure determination for NYSGRC target P097, a hypothetical yeast protein (YNL200C) from chromosome XIV consisting of 246 amino acids and a calculated molecular mass of 27.5 kD. P097 was targeted for structure determination because it is a member of a large protein family (23 sequences currently represented in ProDom, domain PD005835) with unknown biological function. The initial structure was solved by ASDP within 2 h after completing a three-wavelength selenomethionine MAD experiment. Six selenium sites were identified with SOLVE (Terwilliger et al. 1999), using data extending to 2.8 Å resolution. The figure of merit (FOM) for the initial phases was 0.65, which improved to 0.77 after density modification using RESOLVE (Terwilliger et al. 2000b). At this time, the electron density map clearly showed several α-helices and examination of the heavy-atom constellation indicated the presence of a proper twofold noncrystallographic symmetry. The initial phases at 2.8 Å were extended to 1.9 Å with DM/CCP4 resulting in a FOM of 0.86, and this improved map was used for automatic chain tracing with ARP/WARP (Perrakis et al. 1999). The initial model consists of 214 residues in chain A and 217 residues in chain B, or 87% of the structure. Further refinement of the structure was conducted using CNS, crystallography, and NMR system (CNS; Brunger et al. 1998). Several loops exhibited poor electron densities, and several residues were unobserved. The final model, which contained 309 waters had an *R*-value of 0.200 and *R*-free of 0.237 with RMSD in bond distances of 0.006 Å and RMSD in bond angles of 1.2°. There was significant density unaccounted for density near Asn 70 (3.53 Å) and from Gly 142 (3.29 Å). This feature, which likely represents a metal atom, was also close to three water molecules at distances ranging from 3.24 to 3.73 Å.

The P097 structure revealed a three layer α-β-α; sandwich (Fig. 5A), with two molecules forming a tightly packed dimer (Fig. 5B). Each monomer consists of eight β strands and nine α helices. A search using the DALI (Holm and Sander 1993) server showed that P097 is similar to the noncatalytic domain of D-glycerate dehydrogenase (1GDH; Goldberg et al. 1994), with a Z-score of 8.5, a sequence identity of 10% and RMSD of 4.0 Å for Cα atoms. 1GDH has a typical NAD(P)-binding Rossmann fold (Rao and Rossmann 1973), which features at least six β strands, with the first and sixth strand following an alpha helix. The number of β strands in P097 is eight whereas the noncatalytic domain of 1GDH has seven β strands. The conserved regions are shown in Figure 6 that indicate the dimer interface regions and the possible active site regions.

Conclusions

The progress of the NYSGRC after its first year of funding from the NIH indicates that cautious optimism about the overall progress of the structural genomics initiatives is warranted. As of August 31, 2001, the NYSGRC completed 27 X-ray structures that resulted from the examination of more than 500 independent constructs expressed in *E. coli*. Comparative protein structure modeling with these 27 experimentally determined structures produced additional structural information for thousands of protein sequences. These models are publicly available via ModBase (http://guitar.rockefeller.edu/modbase).

With the results of quantitative structure–structure comparisons and homology modeling, we can subdivide our structures into the following four categories (the number of structures and the NYSGRC identifier are shown in parentheses):

1. Structures that represented new protein folds at the time of structure determination and thereby provide novel information about protein sequence/structure space (4/27, P008, P018, P100, T130).

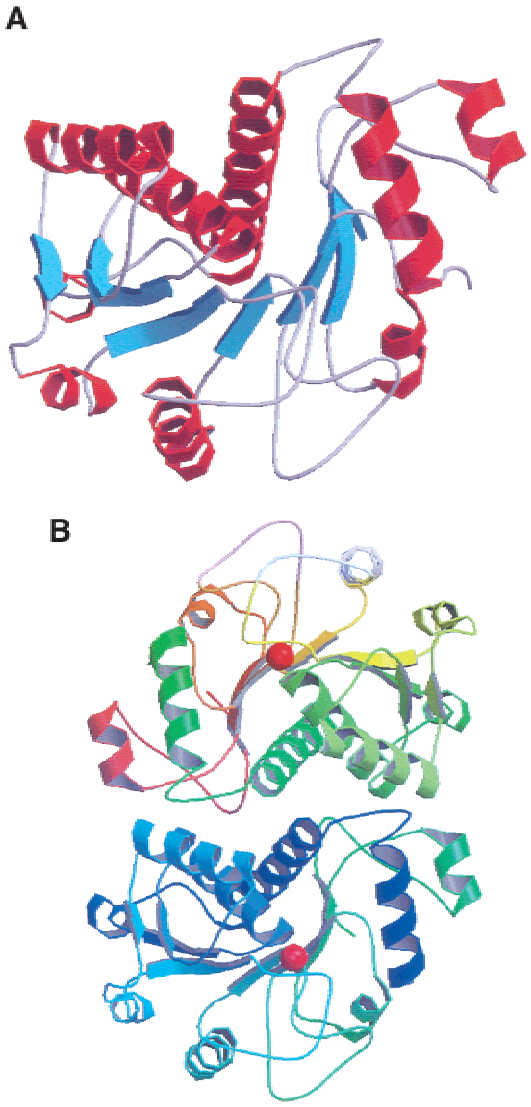

Fig. 5. (*A*) P097 is a three-layer α-β-α sandwich. β Strands are painted as cyan, α helices are painted as red, and loops are painted as gray. (*B*) P097 forms a tightly packed dimer. Large red spheres represent putative metal ions.

2. Structures that are distantly related to previously known protein structures and thereby provide a considerable amount of new information about protein sequence superfamilies (12/27, P007, P096, P097, P109, P111, P111a, T27, T127, T132, T136, T139, T140).
3. Structures that are more closely related to previously known protein structures than class (2) proteins and thereby provide a modest amount of incremental information about subfamilies within protein sequence superfamilies (6/27, P044a, P068, P102, T35, T45, T135).
4. Structures that are very closely related to previously known protein structures and thereby provide little incremental information about protein sequence/structure space (5/27, P003, P048a, T129, T138, B076).

Thus, of our first 27 structures more than half were distantly or entirely unrelated to known structures or folds. Examination of other features of the 27 structures provides en-

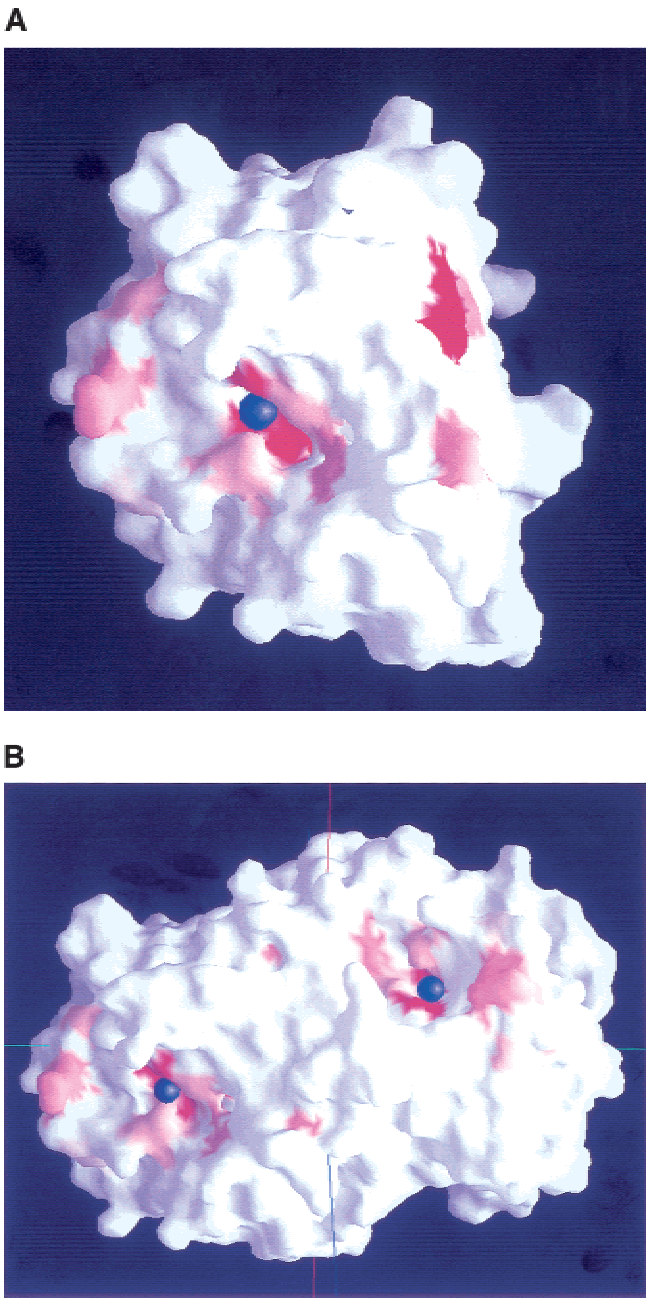

Fig. 6. Multiple alignments of 16 sequences with at least 30% identity to P097 from the Swiss-Prot database revealed several conserved regions (*A*,*B*). (red) Completely conserved residues; (white) nonconserved residues. (*A*) The P097 monomer. Conserved residues at the dimer interface are revealed as a red patch on the righthand side of the molecule. (*B*) The surface of the P097 dimer. The two putative metal ions are represented by blue spheres and may indicate important ligand binding sites as indicated by the conserved residues around the putative metal binding site. Conserved residues at the dimer interface are completely buried.

Table 3. *Summary information for the first 27 proteins crystallized and solved by the NYSGRC*

Range of Organisms	
S. cerevisiae	12
Eubacteria	8
Archaea	3
Human or mouse	4
Methods of structure determination	
MAD/Se	18
MAD/MIR other elements	1
SIRAS	4 (Pt or Hg)
MIR	1
MR from NYSGRC structure	2
Part of larger complex	1
Crystal systems	
Monoclinic	5
Orthorhombic	11
Tetragonal	5
Trigonal or hexagonal	6
Size of the recombinant protein crystallized	280 residues
Resolution limit of the diffraction sample	2.3 Å
Unit cell dimension	86 Å
Longest unit cell dimension	509 Å
Number of residues/asymmetric unit	700
Number of protomers/asymmetric unit	2.5

(MAD) multiple wavelength anomalous dispersion; (MIR) mammalian-wide interpersed repeat; (NYSGRC) New York State Genomics Research Consortium; (Se) selenium; (SIRAS) single isomorphous replacement anomalous scattering.

couragement with respect to the range of targets solved. In Table 3, we show the range of organisms, methods of structure determination, range of crystals types, range of unit cells, average size, and resolution limits of the 27 diffraction samples. A striking feature is that our solved structures are derived from working with recombinant proteins from all three phyla (*Eukarya*, *Archaea*, and *Eubacteria*) and not narrowly distributed evolutionarily. Also, the targets were of significant size, with an average length of 280 residues (14 of the 27 are >250 residues in length). With respect to crystal types and diffraction, large asymmetric units and long unit cell dimensions (including a very challenging case at ~510 Å) and lower symmetry crystal systems (16/27 in monoclinic or orthorhombic) were overcome, and the average resolution limits were 2.3 Å. Although two-thirds of the structures were solved using selenium-MAD, other phasing methods were also critical to productivity. More detailed information on the 27 structures with PDB ID codes and database identifiers is found in Table 4.

Although significant progress is evident in this report, bottlenecks remain in the structural genomics pipeline, particularly in generating sufficient homogeneous protein from a wide array of targets for crystallization trials. Also, as more structures are produced, providing adequate annotation for these structures will become a significant bottleneck. However, this activity cannot be neglected, as it provides the core of the structural genomics effort. In addition, an increasing need for direct functional analysis by biochemical methodologies, not just annotation, is expected to arise, especially for proteins whose function is inferred from structure and where direct experimental tests can be easily imagined. Publication of results is also an issue that has not received proper consideration, and journals may consider short notes or electronic dissemination of annotated structures as a necessary feature of the editorial process beyond the submissions to the PDB required of the structural genomics centers.

Last, some comments are appropriate regarding the impact of structural genomics initiatives on the practice of structural biology in individual laboratories. It is important that structural genomics efforts do not become the only source of structure determination for small, single domain proteins. In many cases, our efforts will not acquire the high-resolution data required for understanding chemical mechanism. Moreover, examination of mutant proteins and substrate or inhibitor complexes is critical in the evaluation of biological and chemical function, and such studies are not part of the mandate of the NIH-funded centers. However, it is expected that there will be some impact, necessitating a focus shift for some laboratories to hard problems, for example, proteins that are difficult to express or structures of multicomponent complexes. The goal of structural genomics is to provide structural models for the biologist, thus permitting improved functional annotation of proteins involved in a wide array of biochemical and cellular processes. This should bring more biologists into the structural fold and promote interest in in-depth structural studies for molecules of biological interest.

Materials and methods

96-Well format for expression and solubility testing of C. elegans constructs

Competent BL21 Star(DE3)pLysS cells (Invitrogen) were prepared using standard protocols and 50 μL aliquots pipetted into each well of a 96-DeepWell plate (Corning or Nunc). Plates were stored at −80°C for later use. Bacteria in DeepWell plates were transformed with the *C. elegans* constructs using standard protocols and grown overnight in 750 μL of Luria–Bertani medium containing 100 μg/mL ampicillin (LB-AMP) at 30°C with vigorous shaking (250 rpm). One hundred fifty microliters of bacterial culture from each well was inoculated into 1.35 mL of prewarmed LB-AMP in a second 96-DeepWell plate. The cultures were grown at 30°C with vigorous shaking to an OD_{600} of 0.5–0.6 (usually 1–3 h of growth). Two hundred fifty microliters of cell culture from each well was pipetted into a standard 96-conical-well plate (preinduction whole cell lysate) and the cells harvested by centrifugation at 2500*g* for 30 min. The supernatant was discarded, and each cell pellet resuspended in 30 μL of SDS-PAGE loading buffer. An additional 300 μL of prewarmed LB-AMP was added to the remaining cultures in the second 96-DeepWell plate, and the cultures

Table 4. *List of solved structure targets from year 1*

PDB ID code	NYSGRC ID code	Organism	Database link	Length (expressed amino acid)	Protein name/other notes
1jd1	P003	*S. cerevisiae*	SP P40037	129	Translation inhibitor YEO7/YjgF family
1jg8	P044a	*T. maritima*	GI 4982322	343	Pyridoxal-5′-phosphate-dependent, L-threonine acetaldehyde-lyase
1jf1	P048a	Human	GI 7513394	205	Transcription negative cofactor 2 α
1fi4	P100	*S. cerevisiae*	SP P32377	396	Mevalonate-5-diphosphate decarboxylase
1i9a	P109a	*E. coli*	GI 6225535	182	Isopentenyl diphosphate:dimethylallyl diphosphate isomerase
1g62	P111	*S. cerevisiae*	SP O12522	245	Translation factor (1F6)
1g61	P111a	*M. jannaschii*	SP O60357	228	Translation factor (1F6)
1cio	P008	*S. cerevisiae*	SP P38075	228	Pyridoxamine 5′-phosphate oxidase
Ihqz	T138	*S. cerevisiae*	GI 113000		Actin-binding protein 1, N-terminal domain (cofilin homolog)
1b54	P007	*S. cerevisiae*	SP P38197	257	Hypothetical protein YBL036C, unknown function, similar to the N-terminal domain of alanine racemase
1jf9	T129	*E. coli*	SP P77444	406	CsdB/NifS, appears to be alpha-family of pyridoxal 5′-phosphate (PLP)-dependent enzymes
1jr7	T130	*E. coli*	SP P76621	311	γ aminobutyric acid metabolism protein
[a]	T136	*M. jannaschii*	SP O58958	153	Methicillin-resistant *S. aureus* (MRSA) auxiliary gene 223 ortholog
[a]	T35	*E. coli*	SP P17113 GI 121191	207	Glucose-inhibited division protein (gid) B
1jss	B076	Mouse	GI 13542894	224	Dietary cholesterol-regulated START protsin
1k47	T27	*S. pneumoniae*	SP Q97SH9	336	Phosphomevalonate kinase (PMK)
[a]	T45	Mouse	GI 2342488	518	Murine collapsing response mediator protein 1 (CRMP1)
1k8f	T139	*S. cerevisiae*	SP P17555	159	Actin-binding domain of yeast cyclase associated protein
1k4z	T140	Human	SP Q01518	157	Adenylyl cyclase-associated protein, human
1f89	P018	*S. cerevisiae*	SP P49954	291	Hypothetical protein YLR351C
[a]	T135	*M. jannaschii*	SP Q58497	438	Bacterial lysine biosynthesis enzyme
[a]	P068	*S. cerevisiae*	SP P40363	299	Hypothetical esterase YJL068C
[a]	P096	*S. cerevisiae*	SP P53889	259	Hypothetical protein YNL168C
1jzt	P097	*S. cerevisiae*	SP P40165	246	Hypothetical protein YNL200C
[a]	P102	*s. cerevisiae*	GI 2198534	491	Glutathione synthase
[a]	T127	Human	GI 9966777	108	Resistin, member of mouse F1ZZ1 family
[a]	T131	*B. subtilis*	SP P20964	429	GTP-binding protein
[a]	T132	*E. coli*	SP P52097	432	ATPase

[a] Final refinement nearly complete; PDB submissions in progress.

were grown for another 20 min at 30°C before induction with 0.5 mM IPTG. Three hours after IPTG addition, 250 μL of cell culture from each well was removed (post-induction whole cell lysate) and processed as described above. The cell cultures in the second 96-DeepWell plate were grown for another hour at 30°C. Well contents were transferred to 1.5-mL microcentrifuge tubes, and cells were harvested by centrifugation at 10,000*g* for 5 min. The supernatants were discarded, and the cell pellets were stored overnight at −80°C. Cell pellets were resuspended in 60 μL Bugbuster solution (Novagen) containing 0.5 M NaCl, 2.5 mM DTT, protease inhibitor cocktail III (Novagen), and benzonase nuclease and rocked at room temperature for 5 min. The lysate was centrifuged at 15,000*g* for 10 min, and the supernatant and pellet fractions were saved for later analysis. Preinduction, postinduction, supernatant and pellet fractions were analyzed by SDS-PAGE.

Isolation, expression, crystallization, and structure determination of Gab protein

The endogenous Gab ortholog from *E. coli* was purified from *E. coli* and crystallized (L.K. Wang and S. Shuman, unpubl.). The Gab DNA coding region was identified and amplified from *E. coli* genomic DNA by PCR, ligated into a non-Topo adapted version of pSUMO, expressed in *E. coli* BL21 DE3 Codon Plus RIL (Stratagene), and purified using Ni-NTA-agarose resin (Qiagen). Gab was further purified by anion exchange (MonoQ 10/10; Pharmacia) and gel filtration (Superdex 200; Pharmacia), eluting from gel filtration with an apparent molecular mass of 120 kD, consistent with an *E. coli* Gab tetramer. Fractions were analyzed by SDS-PAGE, pooled, and concentrated to 10 mg/mL (10 mM Tris-HCl at pH 8.0, 50 mM NaCl, 1 mM DTT).

Ninety-six-well crystallization trials were conducted that produced diffraction quality crystals in several conditions. Gab crystals were grown by hanging drop vapor diffusion against a well solution containing ammonium sulfate and sodium citrate at pH 5.6 to a final size of 0.2 × 0.2 × 0.3 mm. The data presented here were obtained from Gab crystallized in spacegroup I422 (a = b = 126.4 Å, c = 133.6 Å, α = ß = γ = 90°). Diffraction data collection was accomplished with cryopreserved crystals (25% sucrose). Data from native crystals and mercury acetate (HgAc) and thimerosal derivatives were collected at beam line X4A at the National Synchrotron Light Source, processed with DENZO and SCALEPACK (Otwinowski and Minor 1997), and input to SOLVE (Terwilliger and Berendzen 1999), SHARP (La Fortelle and Bricogne 1997), and the CCP4 suite (Collaborative Computational Project 1994) to calculate a 2.0-Å phase set. Density modification was performed with SOLOMON (Abrahams and Leslie 1996). Electron density maps were traced manually with O (Jones et al. 1991; see Table 1A in Appendix), and the model was refined with Refmac (Murshudov 1997). The model contained 321 amino acid residues excluding the N-terminal 14 amino acids that were not included in the expression construct. 2-OG cocrystallization was conducted in anaerobic conditions (argon atmosphere). Coordinates and structure factors are deposited in the Protein Data Bank (accession code 1JR7).

Crystallographic methods for P097 structure determination

The expression and purification of P097 will be reported separately. Selenium-methionine crystals were obtained using purified protein at 1.0 mg/mL in 40 mM Tris, 0.5 M NaCl at pH 6.7 and a crystallization buffer of 18% PEG 3350, 0.1 M Bicine, 0.15 M NH_4formate at pH 8.8. MAD data were collected at beam line X12C at the National Synchrotron Light Source (see Table 2A in Appendix). Data were processed with DENZO and SCALEPACK (Otwinowski and Minor 1997). The spacegroup is $P2_12_12_1$, and the unit cell parameters are 57.75,68.68,125.21, 90.0,90.0,90.0, with diffraction extending to 1.9 Å resolution. Coordinates and structure factors are deposited in the Protein Data Bank (accession code 1JZT).

Acknowledgments

The Gab structure was solved based on data collected from the X4A beamline at the National Synchrotron Light Source, and the support of its staff is acknowledged. Beamline X4A is supported by the Howard Hughes Medical Institute. Please contact C. Lima for further information on Gab. The P097 structure was solved based on data collected from X12C beamline at NSLS and support of its staff is acknowledged. Please contact J. Jiang for further information on P097. This work was supported in part by grants from the Arnold and Mabel Beckman Foundation (C.D.L.), NIH-GM54762 (A.S.), the Mathers Foundation (A.S.), a Merck Genome Research Award (A.S.), NIH-CA81658 (M.V.), and the NIH structural genomics pilot center grant GM62529. Also, support from the Biomedical Technology Resource programs located at the NSLS and funding by the National Center for Research Resources (P41-RR01633 and P41-RR12408) and the Department of Energy are acknowledged. A. Sali and S. Almo are Irma T. Hirschl Trust Career Scientists.

Appendix

Table 1A. *Summary of crystallographic analysis for GAB protein*

	Multiple isomorphous replacement		
	Native (high)	5 mM HgAc	2 mM thimerosal
dMin/λ (Å)	20–2.0/0.9878	20–2.0/0.9878	20–2.0/0.9878
No. of sites	—	2	2
Rsym[a] (%) a overall (outer shell)	7.9 (23.9)	4.9 (15.5)	6.3 (19.8)
Coverage (%) overall (outer shell)	96.9 (94.1)	95.8 (94.9)	74.8 (52.1)
I/σ (I) overall (outer shell)	13.0 (4.2)	9.9 (3.8)	9.1 (4.8)
Reflections (total/unique)	442,770/33,222	446,316/32,810	562,379/25,609
Phasing statistics			
MFID[b] (%)		16.0	28.1
Overall phasing power[c] (centric/acentric)		1.25/1.18	0.74/1.07
Mean FOM[d] (centric/acentric)		0.358/0.371	
Mean FOM[d] after Solomon		0.63	
Refinement			
Resolution range	20–2.0		
No. of reflections >0.0σ	33216		
Total no. atoms/water/Fe(II)	490/320/1		
R[e]/Rfree[f]	0.193/0.240		
Rmsd[g] bond (Å)/angles(°)/B(Å^2)	0.009/1.8/1.63		

[a] Rsym = Σ|I-⟨I⟩|/Σ I, where I = observed intensity, and ⟨I⟩ = average intensity.
[b] MFID (mean fractional isomorphous difference) = Σ||Fph|-|Fp||/Σ|Fp|, where Fp = protein structure factor amplitude and |Fph| = heavy-atom derivative structure factor amplitude.
[c] Phasing power = root mean square (|Fh|/E, where |Fh| = heavy-atom structure factor amplitude and E = residual lack of closure error.
[d] Mean FOM = combined figure of merit.
[e] R based on 95% of the data used in refinement.
[f] R based on 5% of the data withheld for the cross-validation test.
[g] Root mean square deviation of bond lengths, angles, and B factors. Rmsd B is combined for side chain/main chain.
Rc = Σ||Fh(obs)|-|Fh(calc)||/Σ|Fh(obs)| for centric reflections where |Fh(obs)| = observed heavy-atom structure factor amplitude, and |Fh(calc)| = calculated heavy-atom structure factor amplitude.

Table 2A. *Summary of MAD data collection and statistics for P097*

	Peak	Inflection	Remote
Wavelength (Å)	0.9785	0.9788	0.9400
Reflections	39403	39416	39366
R-merge	0.063	0.062	0.062
Completeness (%)	98.3	98.5	98.2
I/sigma >20 (%)	44.7	45.1	45.2
Redundancy >12 (%)	41.6	41.5	44.1

References

Abrahams, J.P. and Leslie, A.G.W. 1996. Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Crystallogr.* **D52:** 30–42.

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 programs. *Nucleic Acids Res.* **25:** 3389–3402.

Baker, D. and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* **5:** 93–96.

Baldwin, J.E. and Abraham, E. 1988. Biosynthesis of penicillins and cephalosporins. *Nat. Prod. Rep.* **5::** 129–145.

Bonnano, J.B., Edo, C., Eswar, N., Pieper, U., Romanowski, M.J., Ilyin, V.A., Gerchman, S.E., Kycia, H., Studier, F.W., Sali, A., and Burley, S.K. Structural genomics of enzymes involved in sterol/isoprenoid biosynthesis. *Proc. Natl. Acad. Sci.* **98:** 12896–12901.

Boulton, S.J., Gartner, A., Reboul, J., Vaglio, P., Dyson, N., Hill, D.E., and Vidal, M. 2002. Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* **295:** 127–131.

Brenner, S.E. 2000. Target selection for structural genomics. *Nat. Struct. Biol.* (Suppl.) **7:** 967–969.

Brenner, S.E., Chothia, C., and Hubbard, T. 1997. Population statistics of protein structures: Lessons from structural classifications. *Curr. Opin. Struct. Biol.* **7:** 369–376.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, N., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., and Warren, G.L. 1998. Crystallography and NMR system (CNS): A new software system for macromolecular structure determination. *Acta Cryst.* **D54:** 905–921.

Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland, T., Lin, D., Sali, A., Studier, F.W., and Swaminathan, S. 1999. Structural genomics: Beyond the Human Genome Project. *Nat. Genet.* **23:** 151–157.

Collaborative Computational Project. 1994. The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr.* **D50:** 760–763.

Davy, A., Bello, P., Thierry-Mieg, N., Vaglio, P., Hitti, J., Doucette-Stamm, L., Thierry-Mieg, D., Reboul, J., Boulton, S., Walhout, A.J., Coux, O., and Vidal, M.A. 2001. Protein–protein interaction map of the *Caenorhabditis elegans* 26S proteasome. *EMBO Rep.* **2:** 821–828.

Evans, S.V. 1993. SETOR: Hardware-lighted three-dimensional solid model representations of macromolecules. *J. Molec. Graph.* **11:** 134–138.

Gaasterland, T. and Oprea, M. 2001. Whole-genome analysis: Annotations and updates. *Curr. Opin. Struct. Biol.* **11:** 377–381.

Gaasterland, T. and Ragan, M.A. 1998. Constructing multigenome views of whole microbial genomes. *Microb. Comp. Genomics***3:** 177–192.

Gaasterland, T. and Sensen, C.W. 1996. MAGPIE: Automated genome interpretation. *Trends Genet.* **12:** 76–78.

Goldberg, J.D., Yoshida, T., and Brick, P. 1994. Crystal structure of a NAD-dependent D-glycerate dehydrogenase at 2.4A resolution. *J. Mol. Biol.* **236:** 1123–1140.

Green, E.D. 2001. Strategies for the systematic sequencing of complex genomes. *Nat. Rev. Genet.* **2:** 573–583.

Greenbaum, D., Luscombe, N.M., Jansen, R., Qian, J., and Gerstein, M. 2001. Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res.* **11:** 1463–1468.

Holm, L. and Sander, C. 1993. Protein structure comparison by alignment of distant matrices. *J. Mol. Biol.* **233:** 123–138.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., and Hood, L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **4:** 929–934.

Jones, T.A., Zou, J.Y., Cowan, S.W., and Kjeldgaard, M. 1991. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr.* **A47:** 110–118.

Kapust, R.B. and Waugh, D.S. 1999. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* **8:** 1668–1674.

La Fortelle, E. and Bricogne, G. 1997. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276:** 472–494.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29:** 291–325.

Mossessova, E. and Lima, C.D. 2000. Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. *Mol. Cell* **5:** 865–876.

Murshudov, G.N., Vagin, A.A., and Dodson, E.J. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr.* **D53:** 240–255.

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.

Nicholls, A., Sharp, K.A., and Honig, B. 1991. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11:** 281–296.

Otwinowski, Z. and Minor, W. 1997. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276:** 307–326.

Perrakis, A., Morris, R.M., and Lamzin, V.S. 1999. Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.* **6:** 458–463.

Pieper, V., Eswar, N., Stuart, A.C., Ilyin, V.A., and Sali, A. 2002. MODBASE: A database of annotated comparative protein structure models. *Nucleic Acids Res.* **30:** 255–259.

Rao, S.T. and Rossmann, M.G. 1973. Comparison of super-secondary structure in proteins. *J. Mol. Biol.* **76:** 241–256.

Roach, P.L., Clifton, I.J., Hensgens, C.M., Shibata, N., Schofield, C.J., Hajdu, J., and Baldwin, J.E. 1997. Structure of isopenicillin N synthase complexed with substrate and the mechanism of penicillin formation. *Nature* **387:** 827–830.

Sali, A. 1998. 100,000 protein structures for the biologist. *Nat. Struct. Biol.* **5:** 1029–1032.

Sali, A. 2001. Target practice. *Nat. Struct. Biol.* **8:** 482–484.

Sali, A. and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234:** 779–815.

Sali, A. and Kuriyan, J. 1999. Challenges at the frontiers of structural biology. *Trends Cell Biol.* **9:** M20–M24.

Sanchez, R. and Sali, A. 1998. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci.* **95:** 13597–13602.

Sanchez, R., Pieper, U., Melo, F., Eswar, N, Marti-Renom, M.A., Madhusudhan, M.S., Mirkovic, N., and Sali, A. 2000. Protein structure modeling for structural genomics. *Nat. Struct. Biol.* (Suppl.) **7:** 986–990.

Schaffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L., and Altschul, S.F. 1999. IMPALA: Matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **15:** 1000–1011.

Shi, W., Ostrov, D., Gerchman, S., Kycia, H., Studier, W., Edstrom, W., Bresnick, A., Ehrlich, J., Blanchard, J., Almo, S., and Chance, M.R. 2002. High-throughput structural biology and proteomics. In *Proteomics: The next phase of genomics discovery*. Marcel Dekker Publishers. (in press).

Terwilliger, T.C. 2000a. Structural genomics in North America. *Nat. Struct. Biol.* (Suppl.) **7:** 935–939.

Terwilliger, T.C. 2000b. Maximum-likelihood density modification. *Acta Crystallogr.* **D56:** 965–972.

Terwilliger, T.C. and Berendzen, J. 1999. Automated MAD and MIR structure solution. *Acta Crystallogr.* **D55:** 849–861.

Valegard, K., van Scheltinga, A.C., Lloyd, M.D., Hara, T., Ramaswamy, S., Perrakis, A., Thompson, A., Lee, H.J., Baldwin, J.E., Schofield, C.J., Hajdu, J., and Andersson, I. 1998. Structure of a cephalosporin synthase. *Nature* **394:** 805–809.

Vidal, M. 2001. A biological atlas of functional maps. *Cell* **104:** 333–339.

Vitkup, D., Melamud, E., Moult, J., and Sander, C. 2001. Completeness in structural genomics. *Nat. Struct. Biol.* **8:** 559–566.

Walhout, A.J., Temple, G.F., Brasch, M.A., Hartley, J.L., Lorson, M.A., van den Heuvel, S., and Vidal, M. 2000a. GATEWAY recombinational cloning: Application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol.* **328:** 575–592.

Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N., and Vidal, M. 2000b. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287:** 116–122.

Walhout, A.J. and Vidal, M. 2001. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods* **24:** 297–306.

Yang, A.S. and Honig, B. 1999. Sequence and structure alignment in comparative modeling using PrISM. *Proteins* **37:** 66–72.

Zhang, Z., Ren, J., Stammers, D.K., Baldwin, J.E., Harlos, K., and Schofield, C.J. 2000. Structural origins of the selectivity of the trifunctional oxygenase clavaminic acid synthase. *Nat. Struct. Biol.* **7:** 127–133.

Web site references

http://asdp.bnl.gov/asda/About/asdp_server.html; expandable computing server.

http://asdp.bnl.gov/asda/ASDP/index.html; including data collection; initial phasing, modeling, refinement, and PDB submission.

http://asdp.bnl.gov; automated structure determination platform for high-throughput structure determination.

http://guitar.rockefeller.edu/modbase; ModBase; relational database of annotated comparative protein structure models.

http://guitar.rockefeller.edu/modweb; ModWeb provides a Web interface to ModPipe and takes as input either a set of sequences or a protein structure.

http://nysgrc.org; New York Structural Genomics Research Consortium.

http://targetdb.pdb.org/; to obtain target lists.

www.genomes.rockefeller.edu/magpie/magpie.html; MAGPIE, to compare organisms at a whole genome level.

www.nigms.nih.gov/funding/psi.html; pooled resources from several institutions.

www.rcsb.org/pdb/strucgen.html; Protein Data Bank Web site.

www.structuralgenomics.org; information from the authors' consortium.