The Southeast Collaboratory for Structural Genomics: A High-Throughput Gene to Structure Factory

MICHAEL W. W. ADAMS,† HARRY A. DAILEY,† LAWRENCE J. DELUCAS,‡ MING LUO,‡ JAMES H. PRESTEGARD,† JOHN P. ROSE,† AND BI-CHENG WANG,*,†

Department of Biochemistry and Molecular Biology, University of Georgia, Athens Georgia 30602, and Department of Microbiology, Center for Biophysical Sciences and Engineering, University of Alabama at Birmingham, Birmingham, Alabama 35294

Received June 3, 2002

ABSTRACT

The Southeast Collaboratory for Structural Genomics consists of four working groups. The protein production group supplies/develops high-output production of *Pyrococcus furiosus, Caenorhabditis elegans,* and selected human proteins. The X-ray crystallography group conducts high-throughput structure production in parallel with production-related research/development in nanocrystallization robotics, capillary crystallization cassette, synchrotron/home X-ray instrumentation, sample mounting robotics, data processing and pipelined structure analysis, combined refinement/validation protocols, and direct use of unlabeled native crystals (Direct Crystallography). The NMR group emphasizes/develops sample screening and backbone structure determination from residual dipolar coupling data. The bioinformatics group implements/develops local database interfaces, pipelined sequence/structure information search/updates, and database/bioinformatics toolkits.

The Southeast Collaboratory for Structural Genomics (SECSG) is a networked Center consisting of five partner institutions in the southeast: the University of Georgia (UGA), the University of Alabama at Birmingham (UAB), the University of Alabama at Huntsville, Georgia State University, and Duke University Medical Center. As one of the original seven NIH-funded Pilot Centers for high-throughput (HTP) structure determination, the Collaboratory aims to develop, integrate, and test all of the constituents for carrying out cost-effective and HTP structural genomics research for both prokaryotic and eukaryotic systems. Genomes under study include *Pyrococcus furiosus* and *Caenorhabditis elegans,* in addition to selected human genes.

To accomplish these objectives, the Collaboratory has been structured into one administration unit and four working groups: protein production, X-ray crystallography, NMR spectroscopy, and bioinformatics, described below.

SECSG Protein Production

***Pyrococcus furiosus* Proteins.** The goal of this project is to develop high-throughput (HTP) gene cloning, expression, and protein purification protocols that will ultimately produce active, recombinant versions of virtually any gene

* Corresponding author. E-mail: wang@BCL1.bmb.uga.edu.
† University of Georgia.
‡ University of Alabama at Birmingham.

Michael W. W. Adams received his BS (1976) and PhD (1979) in Biochemistry from the University of London, England. He joined the Biochemistry and Molecular Biology (BMB) Department at the University of Georgia (UGA) as an Assistant Professor (1987), following six years as a Research/Senior Biochemist at the Corporate Research Laboratories of Exxon Research and Engineering Co. in Annandale, New Jersey, and two years as a post-doctorate Research Associate at Purdue University. He is currently Distinguished Professor of Biochemistry, Molecular Biology and Microbiology. His research interests involve structural and functional genomics of hyperthermophilic organisms that grow near 100 °C.

Harry A. Dailey received a BA (1972) in Bacteriology and PhD (1976) in Microbiology from UCLA. He joined the Department of Microbiology at UGA(1980) as an Assistant Professor following postdoctoral studies at the University of Connecticut Health Center. He served as Head of the Department of Microbiology at UGA (1987−1996) and became jointly listed as Professor of Microbiology and BMB (1996). He was appointed the first Director of the Biomedical and Health Sciences Institute at UGA (2001). His research interests involve heme biosynthesis, particularly the structure/function and regulatory aspects of the terminal two pathway enzymes and their relation to the human genetic diseases named porphyrias.

Larry J. DeLucas received his BS (1972) and MS (1974) in Chemistry from University of Alabama Birmingham (UAB). He also has a BS in Physiological Optics (UAB, 1979), an OD in Optometry (UAB, 1981), and a Ph. D. in Biochemistry (UAB, 1982), and he has held several positions at UAB. In 1994, he became Director of both the Center for Biophysical Sciences and Engineering and the Comprehensive Cancer Center X-ray Core Facility. His research interests are microgravity crystallization and nanocrystallization robotics.

Ming Luo received his BS of Chemistry from Wuhan University, China (1982), and his Ph.D. of Biological Sciences from Purdue University (1987). He is a currently a Professor at UAB (since1987) and the Associate Director of the Center for Biophysical Sciences and Engineering at UAB. His other research interests include the 3D structure of viruses and drug design. He is also a Chang Jiang Professor (adjunct) at Peking University, China.

James H. Prestegard received his PhD in chemistry from Cal Tech (1971). He moved directly to a faculty position in the Chemistry Department of Yale University where he stayed until 1997, when he moved to UGA as Professor and Eminent Scholar in Nuclear Magnetic Resonance Spectroscopy. He conducts research in the Complex Carbohydrate Research Center and contributes to teaching in the Chemistry and Biochemistry Departments. Research interests center on structural characterization of biomolecules including proteins, carbohydrates, and lipids.

John P. Rose received his BS Chemistry from Benedictine College, Atchison, KS (1975), and his PhD in Physical Chemistry from Rutgers University, Newark, NJ (1980). After 15 years at University of Pittsburgh (Department of Crystallography), he relocated to UGA (1995). He is currently a Senior Research Scientist, BMB, UGA, and the Assistant Director of the Southeast Regional Collaborative Access Team, Advanced Photon Source, Argonne National Laboratory. His research interests include X-ray instrumentation, automation and highly accurate methods of X-ray data collection.

Bi-Cheng Wang received a BSc in Chemical Engineering (1960) from Cheng Kung University, Taiwan, and a PhD in Chemistry at the University of Arkansas (1968). After a postdoctoral appointment at Cal Tech (1968−70), he spent the next 25 years at the VA Medical Center, Pittsburgh, and the University of Pittsburgh. In 1995, he joined UGA as Professor and Eminent Scholar in Structural Biology (Crystallography). In 1997, he became Director of the Southeast Regional Collaborative Access Team for the access of synchrotron at the Advance Photon Source, Argonne National Laboratory, and, in 2000, the Director and PI of the Southeast Collaboratory for Structural Genomics. His current research interests are structure−function of biologically relevant macromolecules and development of high-throughput crystallographic methods, including sulfur phasing and determination of protein crystal structures directly using native crystals.

10.1021/ar0101382 CCC: $25.00
Published on Web 02/22/2003

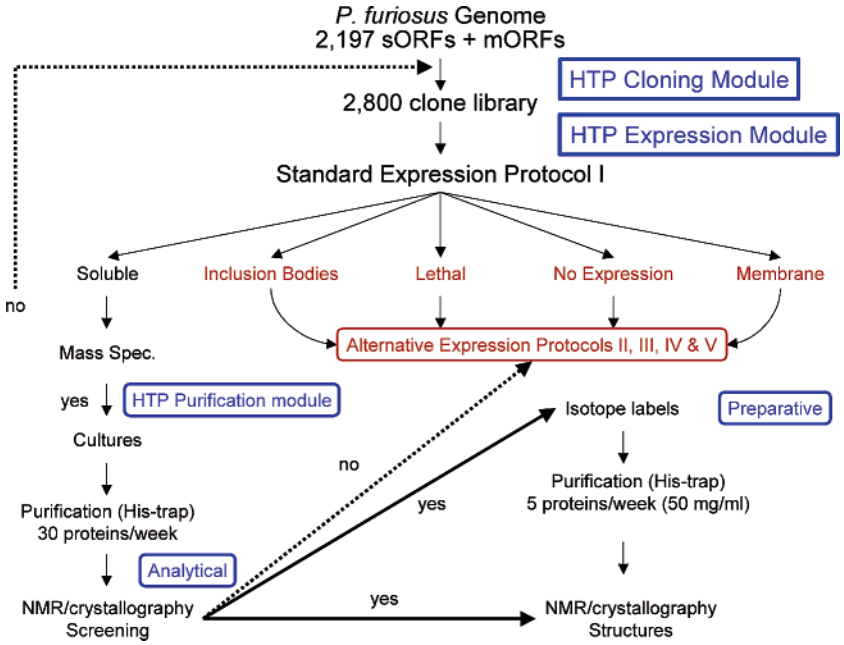

FIGURE 1. A flowchart of the Pf protein production.

or set of genes regardless of the source organism. A major problem with single open-reading frame (ORF), robotic-based expression systems is that they intrinsically select for cytoplasmic, homomeric, unmodified, and/or co-factorless recombinant proteins. Moreover, the overall success rate in producing proteins in their native functional states is probably less than 20%. Clearly, not all ORFs in any genome are created equal from an expression perspective, and there is not a universal expression system that is able to produce recombinant versions of any protein. Our objectives are to develop a multifaceted HTP expression and protein purification system that will also accommodate genes encoding proteins that are multi-subunit, membrane-bound, and/or contain complex co-factors, in addition to homomeric, cofactorless proteins.

To develop this technology, the genome of the prokaryote, *Pyrococcus furiosus* (Pf), is being used as a model system. Pf was chosen because of its evolutionary position as a slowly evolving archaeon (archaebacterium) and because it has a relatively small and well-defined genome. This is 1.9 Mb in size and contains approximately 2200 ORFs.[1] The objective is to clone and express all of these ORFs, with no a priori target selection. It is thought that the genome contains representatives of most, if not all, protein families. As is typical for every genome so far sequenced, about half of the ORFs in the Pf genome encode ORFs of completely unknown function.[2] About 700 of the ORFs are predicted to be organized in operons, suggesting that they encode either multisubunit complexes or include accessory proteins for assembly of the active enzyme. Using the HTP-system with multiple expression systems in various hosts, all 2200 single ORFs (sORFs) will be expressed individually and, where genome analyses indicate, as multiple ORFs (mORFs) to yield all proteins, from the simplest, cytoplasmic protein to the most complex membrane protein assembly.

In the first stage of the project, the approximately 3000 sORFs and mORFs are being expressed in a conventional *Escherichia coli* system (Figure 1) using simple variations in conditions (temperature, strain, inducer concentration, etc.) to optimize expression of each ORF. All proteins are produced with a hexaHis tag at the N-terminus. This is used to ultimately facilitate purification, but the tag is also being used, via an immunological assay, to determine the extent and cellular location of the expressed protein, e.g., in the soluble fraction, as an inclusion body, or membrane-associated. Such analyses are carried out in a multiple 96-well format. Once expression conditions have been optimized on the small scale, cultures are grown on the 1 L scale so that sufficient protein (currently 0.5 mL of >0.3 mM) can be produced for crystallographic and NMR screening. The protein in such samples is at least 95% pure by SDS-gel analysis, is of the correct mass (± 0.01%) as determined by mass spectrometry, and its UV−visible absorption spectrum and metal content (using plasma emission spectroscopy) are also determined. To date (January 2003) a total of 258 different sORFs have been expressed on a large scale, and 175 proteins have been purified. So far, 106 of these have passed through the various analytical procedures and have been supplied to the crystallography group for screening, with 81 of them (76%) yielding crystals.

Overall we expect about a 20% success rate in obtaining stable (nonprecipitating), soluble protein when all of the ORFs in the Pf genome have been passed through the standard *E. coli* expression system. The remaining 80% or so will then be obtained using various prokaryotic and eukaryotic expression systems, yet to be fully developed, that facilitate the production of membrane proteins and proteins with complex cofactors. Eventually, using the results from the Pf system, protocols will then be developed whereby the successful expression of any given gene

Table 1. Estimate of Number of Crystals Produced Screening 350 Genes /Month

success = $N \times R$
N = number of experiments
R = success rate
starting with 350 genes/month
$N_{expressed} = 350 \times R$ (50%)
$N_{soluble} = 175 \times R$ (35%)
$N_{purified} = 61 \times R$ (50%)
$N_{crystallized} = 30 \times R$ (40%) = 12
one year total = $12 \times 12 = 144$

can be predicted on the basis of sequence and, where available, genomic analyses. Predictions will be tested using known and unknown ORFs represented in other prokaryotic genomes and in eukaryotic cDNA libraries. In the longer term, methodologies will be incorporated into the HTP system to accommodate various protein modifications commonly employed in eukaryotic systems.

***Caenorhabditis elegans* Proteins**. The sequencing of the *Caenorhabditis elegans* genome was completed in 1998,[3] and it was predicted that the *C. elegans* genome encodes at least 19 000 genes. Our ultimate goal is to determine as many as possible three-dimensional structures of *C. elegans* proteins using HTP crystallography.

It is evident that expression of soluble proteins in bacteria such as *E. coli* is more difficult for eukaryotic genes. In our pilot study, we designed a process that could produce 100 crystal structures per year when fully operational. As shown in Table 1, we need to process about 350 genes every month if a certain success rate is assumed for each step. The lowest success rate is at the step of purification of soluble proteins. It appears that we could only expect to produce soluble proteins for about 10% of the genes we start with if no prior selection is performed to optimize the outcome.

The cDNA of prospective *C. elegans* genes was cloned into the ENTRY vector of the GATEWAY system[4] (Invitrogen) and stored in 96-well plates. The insert in the ENTRY vector was then transferred into an expression vector, pET15b-DEST, which is a GATEWAY-compatible vector made from pET15b (Novagen), via the LR reaction in 96-well plates on a robot. Using a QIAGEN BioRobot 9600 and 96-well Turbo miniprep kits, the expression plasmids were amplified and purified. The resulting plasmids were transformed into *E.coli* BL21(DE3) strain by heat-shock on a robot. The transformed bacteria were cultured in 1 mL volume to a density of 0.5 OD_{600nm} at both 18 °C (to determine total protein) and at 35 °C (to screen for soluble proteins). After soluble expression of a protein was identified, 6 L of culture was prepared, and the protein was purified by a Ni-affinity column, a gel filtration column (S-75, Pharmacia) and an ion exchange column (Resource, Pharmacia) before being subject to crystal screening.

By this procedure, we were able to achieve the expected success rate at each step and obtained (January 2003) over 248 soluble proteins from about 1529 genes with 27 proteins crystallized so far. However, further modification was required in order to obtain purified proteins that could be crystallized. For example, protein F53F4.3 was

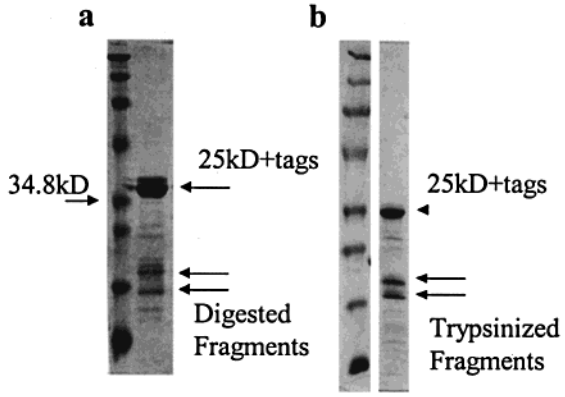

FIGURE 2. SDS−PAGE of the *C. elegans* F53F4.3 protein. (a) Protein digestion observed after storage and (b) similar fragments produced by limited trypsin digestion.

predicted to have a molecular weight of 25.2 kD. This protein was expressed as soluble in *E. coli*, and about 50 mg of purified protein could be obtained from 6L of culture. However, this protein could not be crystallized in this form after screening over 600 crystallization conditions. After storage of purified F53F4.3 protein, it was discovered that the protein was digested to smaller fragments (Figure 2a). Similar fragments could also be produced by limited trypsin digestion (Figure 2b). The fragments were identified by mass spectrometry and N-terminal sequencing to correspond to a C-terminal domain (101−229), a conserved domain (CAP-Gly) that binds tubulin.[5] This fragment was then subcloned as a separate protein in the pET15b vector. The purified C-fragment was easily crystallized (screened with 100 conditions), and its structure was determined.[6] This experience demonstrated that subcloning of digested fragments from soluble protein preparations is inevitable in order to identify the portion of a gene product that would ultimately be crystallized.

Human and Other Eukaryotic Proteins. The eukaryotic protein production group at UGA has approached the problem of protein expression by utilizing this group's past experiences. Briefly stated, we have found that a number of diverse prokaryotic and eukaryotic proteins are not produced at acceptable levels as soluble, nonaggregated proteins with correct cofactor insertion when the T7 promoter system is employed. While the T7 system works well for a number of proteins, it frequently produces inclusion bodies which may be due to the rapid rate at which proteins are synthesized once the inducer is added to the culture medium. A number of other laboratories have also experienced this problem and have addressed it by either lowering growth temperatures to as low as 18 °C in order to slow protein production, or denaturing the isolated inclusion bodies and then attempting refolding of the protein. Neither of these solutions is ideal for high throughput. Lowered growth temperature results in longer expression times and refolding entails an additional step with limited success.

Our group has found that the Trc (Tac) promoter, while much weaker than the T7 promoter, is more than satisfactory to produce the level of protein production needed

for structural studies. The system developed by our group uses the pTrcHis plasmid (Invitrogen) and clones the cDNA whenever possible into the NheI site on the 5′ end so as to give a product protein with an amino terminal his_6 tag and only six other residues. We have observed that omitting the enterokinase cleavage site and additional residues from the parent plasmid facilitates crystallization, while inclusion of these extra amino acids can prevent crystallization. The expression plasmid is transformed into *E. coli* JM109. The specific protocol employed does not use IPTG as an inducer but involves growing cells into stationary phase, at which time the Tac promoter is induced by tryptophan depletion in the medium. Two distinct advantages to this growth protocol are that it is not necessary to monitor culture OD, which saves time, and that it is not necessary to add inducers such as IPTG, which saves money. This protocol yields 1−20 mg protein per L for cDNAs that are expressed in *E. coli*. Our protocol also deviates from most high-throughput methods in that we do not utilize a 96 well screening format but directly try 1 L expression of all vectors. The rationale for this approach is that scale-up from a 1−5 mL screening culture to a 1 L format is not always successful due to the variation between culture conditions in 1 mL versus 1 L volume vessels. In our system failed expressers are cataloged and will be screened using different expression protocols at a later date.

To date we have employed this system with a variety of cDNAs and have successfully produced both soluble and membrane-associated proteins with a variety of cofactors, including FAD, pyridoxal phosphate, iron−sulfur clusters, various metals, heme, and the unique cofactor dipyrrole methane. Proteins ranging in size from under 10 to over 50 kD have been produced, and a number of these are homodimeric or tetrameric.

During our initial trials we utilized one enzyme, ferrochelatase, from a variety of sources to examine factors which may affect protein production. This enzyme carries out the same reaction in all organisms, but its physical characteristics vary considerably. In some organisms it is a monomer; in others it is a dimer. In some it contains a [2Fe−2S] cluster, while in others it has no cofactor. Depending on the source organism, the enzyme may be either soluble or membrane-associated. When these various ferrochelatases from prokaryotes and eukaryotes are expressed in the system described above, one finds that protein expression varies from about 20 mg/L culture to less than 1 mg/L. The level of protein productivity does not appear to correlate with the presence or absence of cofactor, predicted cellular location or subunit composition; nor is there a correlation of expression levels with cDNA GC content. Expression in systems that provided increased levels of rare codon tRNAs did not increase protein production.

SECSG X-Ray Crystallography

HTP Crystal Screening. The *NanoScreen* system is a novel technology for automated and efficient screening of crystallization conditions, which greatly reduces the amount of protein necessary. Thus, the number of proteins that can be investigated will be increased. The *NanoScreen* (Figure 3) can accurately dispense as little as 10 nL of solution in each experiment. Several thousand experiments can be performed daily.

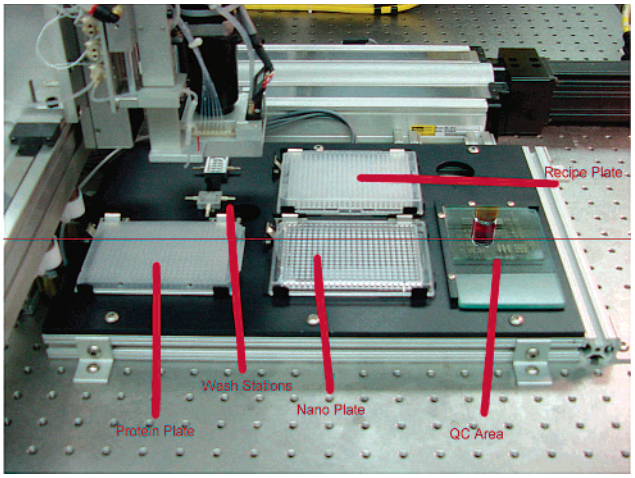

FIGURE 3. Photographs of the *NanoScreen* HTP crystallization system under development at SECSG showing (from left to right) the protein storage plate, wash station, nanocrystallization plate, recipe plate, and quality control station. *NanoScreen* uses the microbatch under oil technique and is capable of dispensing 10 nL drops of solutions with varying viscosity.

One strategy the *NanoScreen* system will be using is incomplete factorial experimental design. Experimental conditions are chosen to reflect a statistically relevant sample of all possible chemical conditions. This strategy utilizes a subset of all possible experimental/solution combinations to discover efficient crystallization conditions. The initial experimental screen is based on an incomplete factorial matrix that encodes the values of experimental variables at several different levels. These levels are distributed evenly with respect to each other using a statistically driven computer program to give a balanced design which allows the extraction of more information than can be obtained from the sparse matrix or random designs used in other systems.

A novel capillary crystallization cassette system has also been developed.[7] This system combines the crystallization and crystal mounting steps eliminating the need for crystal transfer and is advantageous for handling sensitive and/or fragile crystals. The device provides the capability of data collection at both cryogenic and ambient temperatures. This feature is useful for crystals that cannot be flash-cooled in the conventional loop-mounting practice.

HTP Structure Determination. The X-ray Crystallography group is carrying out HTP structure production in parallel with production-related research/development, such as sample-mounting robotics, synchrotron/home X-ray instrumentation including Cr X-rays, data processing and pipelined structure analysis, combined refinement/validation protocols, and direct use of native crystals in a process we term “Direct Crystallography”. The goal of Direct Crystallography is to reduce the number of inter-

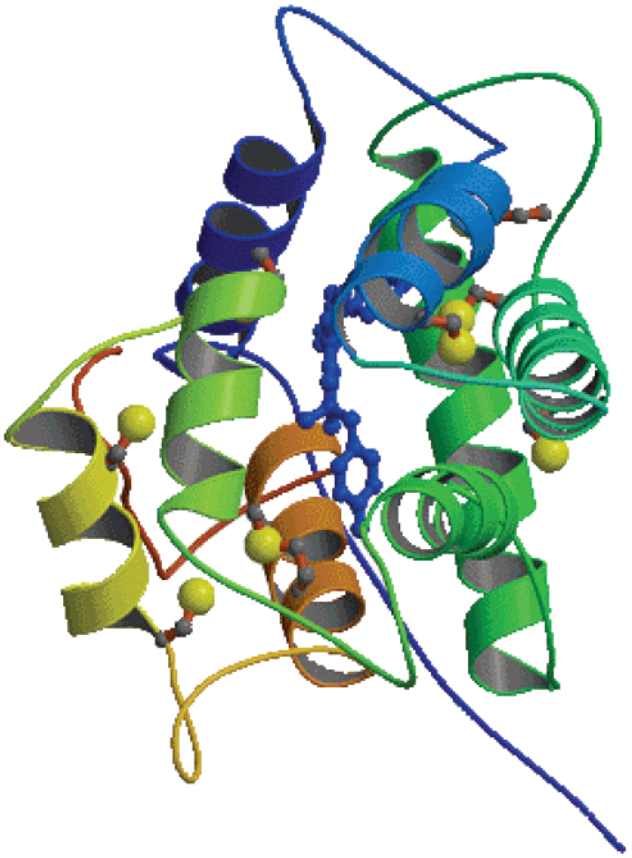

FIGURE 4. The crystal structure of the photoprotein obelin determined from its sulfur substructure. SAS data were collected at 17ID IMCA-CAT, APS using 1.74 Å X-rays. Eight sulfur atoms were located using SOLVE. The structure was solved using ISAS (*Methods Enzymol.* **1985**, *115*, 90).

mediate steps in the structure determination process:

$$\text{Gene}-\text{Protein} \xrightarrow{\text{Direct Crystallography}} \text{Structure}$$

Unlike traditional structure determination methods such as multiwavelength anomalous dispersion (MAD), Direct Crystallography aims to use native crystals and a single set of single-wavelength anomalous scattering (SAS) data. This saves the time and expense of preparing selenomethionyl derivatives. Direct Crystallography exploits the anomalous scattering signal of metals and/or sulfur naturally present in the crystal. This is advantageous since about 30% of proteins are known/estimated to be metalloproteins and almost all proteins contain sulfur. For crystals containing metals, we consider them as Direct Crystallography's "low-hanging fruits", since their structures should be easily obtained. However, the anomalous scattering signal of sulfur alone is relatively weak ($\Delta f'' = 0.124-1.142$ for $\lambda = 0.71-2.29$ Å). For those crystals that do not contain metal (Direct Crystallography's "high hanging fruits (HHF)") phasing, using the anomalous scattering signal of sulfur atoms becomes significantly more difficult. We are actively developing techniques to accurately record, process, and use these weak signals for HTP structure determination. Our successful examples of this approach include *O. longissima* obelin[8] (Figure 4) by S(Sulfur)-SAS, human ferrochelatase[9] by Fe−SAS, yeast sc-MTFB[10] by Xe−SAS, *C. elegans* F53F4.3[6] by S−SAS, *P. furiosus* ORF 65527 by X-SAS (where X is unknown element(s), see below), and *P. aeruginosa* Lectin-1 by S−SAS.

The Pf 65527 and Pa lectin-1 studies illustrate the potential of the Direct Crystallography approach. The Pf 65527 sequence shows no metal binding motifs; thus, the initial experiments were designed for S−SAS analysis. However, the Bijvoet difference Patterson map showed numerous peaks much stronger than those expected for sulfur, and Harker analysis gave nine heavy-atom sites. These sites were arbitrarily assigned as iron for phasing purposes. The resulting electron density map yielded the complete structure, which has been refined,[11] using data to 1.97 Å, giving an *R*-value of 22.9% ($R_{free} = 27.2\%$). Thus, the exact knowledge of the anomalous scattering content of the crystal may not be required to produce interpretable electron density maps.

Initial attempts to solve the structure of Pa lectin-1 using "in-house" sulfur−SAS data (Rigaku FRD/MaxFlux optics) failed. However, using the combination of 2.5 Å resolution data collected with 1.74 Å X-rays (17ID IMCA-CAT, APS) and 1.5 Å resolution data collected using 1.0 Å X-rays (5.03 Advanced Light Source), the Pa lectin-1 structure could be solved and automatically traced without user intervention. Next, a pipelined version of the SOLVE (2.02)[12]/wARP (5.1)[13] approach used on the synchrotron data was applied to the home source data in a procedure that we term "SCA2STRUCTURE". The pipelined procedure uses the computing cluster's 128 processors to spawn hundreds of SOLVE/wARP jobs using various combinations of input parameters. SCA2STRUCTURE was successful at automatically determining the P1-Lectin structure. The structure built in this manner had an *R* value of 19.8% with an R_{free} of 25.6%. Thus, it appears that fine-tuning program input parameters via a pipelined approach can yield structures in cases where traditional treatments fail.

We have recently installed a chromium confocal optics system (Rigaku/MSC) on an in-house chromium X-ray source and are designing experiments to test the feasibility of using this source to determine protein structures. The use of chromium radiation ($\lambda = 2.29$ Å) is attractive for sulfur SAS experiments since the sulfur anomalous scattering signal is doubled ($\Delta f'' = 1.142$) when compared with Cu Kα X-rays ($\Delta f'' = 0.557$). Initial results show that this approach is promising and more tests are underway.

In addition, SECSG is also collaborating with SER-CAT (Southeast Regional Collaborative Access Team), sector 22 Advanced Photon Source on automated sample mounting and recovery, remote data collection, telepresence, and intelligent data collection and processing. The goals of this collaboration are to (1) recover beam time lost due to safety considerations and user fatigue, (2) reduce the personnel and operating costs, and (3) collect only the data needed to produce the structure. The SER-CAT collaboration has already let to a commercially available crystal-mounting robot ACTOR (market by Rigaku/MSC, loosely based on UGA/SER-CAT/OSS initial designs) and

small format motorized X−Y crystal centering device required for automated crystal mounting that has been adopted by the APS Structural Biology Center (sector 19) for its beamlines.

SECSG HTP NMR

NMR is not normally considered a high-throughput structural technique, as traditional methods for NMR determination of complete proteins structures can require 4−6 weeks for data acquisition and subsequent months for assignment and structure calculations.[14] In some cases, where structures promise to be particularly unique and are not likely to be attainable by any other methods, this investment in time may be justified. However, in general exploration of ways that NMR can complement other structural approaches seems more productive. Our NMR core attempts to do this through several applications: preliminary screening of samples as an aid to selection of the best structure determination methodology, rapid backbone structure determination of small proteins which are unlikely to crystallize, and validation of predicted or modeled structures of larger proteins that are unlikely to crystallize.

Simple one-dimensional NMR spectra can be used as a preliminary screening device to predict whether a protein is likely to crystallize. There is substantial anecdotal evidence that proteins with segments having low structural order are difficult to crystallize;[15] our approach is based on an ability to detect low structural order from one-dimensional NMR spectra. A simple magnetization transfer experiment that begins with water magnetization can highlight amides in disordered regions of proteins.[16] To date, only 19 proteins have been both screened by NMR and subjected to crystallization trials, but of the 10 classified as well-structured, all but 2 have crystallized, and of the 9 classified as partially unstructured, only 5 have crystallized. The correlation is encouraging, and we expect this to lead to a robust method of targeting NMR structural determination to those proteins least likely to crystallize.

For structural NMR applications, one must still work to reduce the time required. Our efforts have gone into developing methodology for the determination of backbone structures rather than complete structures. This is not because complete high-resolution structures are undesirable, but because a number of goals of the structural genomics initiative, including providing a better database representing fold families, can be met with structures of backbones alone. The methods are heavily dependent on residual dipolar coupling data, which can be acquired with far greater efficiency than traditional NOE data.[17] We recently reported the successful determination of the structure of rubredoxin as a test case. This is a small 54-amino-acid protein that is highly soluble. It allowed the development of a rapid method of simultaneous resonance assignment and backbone structure determination relying only on dipolar couplings and chemical shift values.[18] Figure 5 shows a recently refined

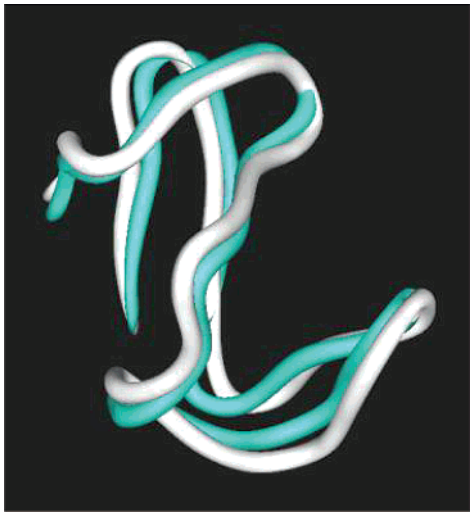

FIGURE 5. Superimposed NMR structure (gray) and Xray structure (cyan) (Koradi, R.; Billeter, M.; Wuthrich, K. *J. Mol. Graphics* **1996**, *14*, 51). The NMR structure is based on the collection of residual dipolar coupling data that yield backbone orientation constraints that does not require prior assignment of resonances. The acquisition of data required about 10% of time needed for traditional NOE-based approaches to protein structure determination.

version of the structure that agrees within a 1.64 Å rmsd to the backbone atoms of a homologous high-resolution X-ray structure (1BRF). Efforts are currently underway to apply this approach to larger (10−20 kDa) proteins that are on our structural genomics target list.

Proteins larger than those amenable to the above methods are not necessarily beyond the reach of NMR investigation. Through a simple analysis of the distribution of residual dipolar couplings, NMR can be used as a rapid method for validating and refining structures of larger proteins (20−80 kDa). These structures may either be predicted on the basis of sequence homology or ab initio structure prediction.[19,20] NH vectors are extremely sensitive to their orientation relative to the ordering frame and thus provide a great deal of structural information. In this approach there is no need to assign resonances, as the entire distribution of NH dipolar couplings can be compared with those calculated from the predicted structure. Preliminary applications have demonstrated the ability of this method to recognize homology of proteins chosen from different fold families.[21]

The approaches outlined above allow us to remain optimistic about the role of NMR in structural genomics. It can provide structures for proteins that may not easily crystallize, and it can do so rapidly, especially if a high-resolution backbone structure is the primary objective. We are currently expanding and refining our methods to make them applicable to a large number of protein targets.

SECSG Bioinformatics

The goals of the bioinformatics team are to build/maintain appropriate local databases and to help transforming massively heterogeneous data into a human-comprehensible form to facilitate the study of biological problems. To achieve these objectives, the SECSG team is actively

(1) developing bioinformatics software/databases for tracking experimental results, (2) streamlining the integration/correlation of information from disparate sources, (3) regularly updating important databases related to our selected targets, (4) disseminating information to the research community, and (5) assisting in developing high-throughput pipelined approaches for analyzing biological information, predicting/modeling protein structures, and validating the determined structures.

To maximally use the specific expertise within each group, the individual groups are developing/maintaining their own Laboratory Information Management System (LIMS) and provide database interfaces in the form of Common Gateway Interface (CGI) query strings, Excel spreadsheets, tab-delimited tables and Extensible Markup Language (XML) format files. The bioinformatics team in turn develops generic parsers for these database interfaces and for converting different format files into an XML based exchange layer, which is independent of operating system/databases. Our electronic progress report (http://www.secsg.org/cgi-bin/report.pl) was designed based on the above strategy. This report is generated weekly by integrating information from the working groups (protein production, X-ray crystallography, and NMR). Thus, researchers can cross check progress in the individual groups.

Protein sequence and structure information are generated at a rapid rate. It is important to know what newly characterized proteins or deposited structures are related to the protein or protein homologue we are studying. We have automated the search process and are using PSI-Blast[22] to search nonredundant sequence databases monthly and using Blast2 and a refined PSI−Blast procedure to search PDB sequences weekly. The search hits are integrated into the progress report.

We are also building infrastructure for correlating biological information from disparate sources providing toolkits based on mathematical set and combinatorial graph techniques. One Perl (http://www.perl.org) based toolkit that uses bioperl (http://bio.perl.org) has already been applied in analyzing neighbor gene relatedness for complete bacterial genomes released at NCBI. It will be extended to support future correlation of genes with other genomic, structural and functional information. We are also seeking collaborations with different computational biology groups to employ their protein structure prediction and modeling pipeline locally in order to closely connect to our high-throughput experiments.

To facilitate the export of information, a structure deposition and quality control protocol is under development. The protein and experimental information from the structure determination is extracted from the SECSG databases and integrated/transformed into the format acceptable by the Protein Data Bank[23] (PDB). Our collaboration with the Richardson group[24] will ensure the deposition of the highest-quality structural information.

With the IBM SUR award, a 64-node dual processor Linux Cluster and associated hardware, we have implemented sequence comparison, pattern discovery, and crystallographic structure solving pipeline in a high-throughput environment. Within the first 2 weeks of the implementation of the SCA2STRUCTURE pipeline, we produced the first automatically determined structure in SECSG. In addition, we are developing visualization tools to help mine large datasets generated by high-throughput computation.

Funding for this project is being provided by the National Institute for General Medical Sciences (P50-GM62407), The University of Georgia, The Georgia Research Alliance, and the University of Alabama, Birmingham. The authors wish to thank Phil Brereton, Mike Carson, Lirong Chen, Tamara Dailey, Jim Finley, Zheng-Qing Fu, Alan Gingle, Robert Harrison, Michi Izumi, Francis E. Jenney. Jr., Peter LeBlond, Songlin Li, Chi-Hao Luan, Zhi-Jie Liu, Dawei Lin, Xinli Lin, Jonathan Myers, Kristen Mayer, Mike Mayer, Edward Meehan, Lisa Nagy, M. Gary Newton, Joseph D. Ng, Farris Poole, Jeremy Praissman, Shihong Qiu, David Richardson, Jane Richardson, Florian Schubot, Ashit Shah, Claudia Shah, Andrea Steven, Frank Sugar, Wolfram Temple, Homay Valafar, and Irene Weber for their various contributions to the work described in this article. A special thanks to Gary Newton for his assistance in assembling/editing the manuscript. Data were collected at beamline 17-ID (Industrial Macromolecular Crystallography Association Collaborative Access Team (IMCA-CAT)) operated under a contract with Illinois Institute of Technology at the Advanced Photon Source (APS). Use of the APS was supported by the U.S. Department of Energy, Basic Energy Sciences, Office of Energy Research, Contract No. W-31-109-Eng-38.

References

(1) Weiss, R. B. Genbank direct submission, 2002.
(2) Robb, F. T.; Maeder, D. L.; Brown, J. R.; DiRuggiero, J.; Stump, M. D.; Yeh, R. K.; Weiss, R. B.; Dunn, D. M. Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: Implications for physiology and enzymology. *Methods Enzymol. 330*, 134−157.
(3) Anonymous. Genome sequence of the nematode C-elegans: A platform for investigating biology. *Science* **1998**, *282*, 2012−2018.
(4) Walhout, A. J. M.; Temple, G. F.; Brasch, M. A.; Hartley, J. L.; Lorson, M. A.; van den Heuvel, S.; Vidal, M. GATEWAY recombinational cloning: Application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol. 328*, 575−592.
(5) Riehemann, K.; Sorg, C. Sequence Homologies between 4 Cytoskeleton-Associated Proteins. *Trends Biochem. Sci.* **1993**, *18*, 82−83.
(6) Gavira, J. A.; Toh, D.; Lopéz-Jaramillo, J.; García-Ruíz, J. M.; Ng, J. D. Ab initio crystallographic structure determination of insulin from protein to electron density without crystal handling. *Acta Crystallogr.* **2002**, *D58*, 1147−1154.
(7) Li, S. L.; Finley, J.; Liu, Z. J.; Qiu, S. H.; Chen, H. L.; Luan, CH.; Carson, M.; Tsao, J.; Johnson, D.; Lin, G. D.; Zhao, J.; Thomas, W.; Nagy, L. A.; Sha, B. D.; DeLucas, L. J.; Wang, B. C.; Luo, M. Crystal structure of the cytoskeleton-associated protein glycine-rich (CAP-Gly) domain. *J. Biol. Chem.* **2002**, *277*, 48596−48601
(8) Liu, Z. J.; Vysotski, E. S.; Chen, C. J.; Rose, J. P.; Lee, J.; Wang, B. C. Structure of the Ca^{2+}-regulated photoprotein obelin at 1.7 angstrom resolution determined directly from its sulfur substructure. *Protein Sci.* **2000**, *9*, 2085−2093.
(9) Wu, C. K.; Dailey, H. A.; Rose, J. P.; Burden, A.; Sellers, V. M.; Wang, B. C. The 2.0 angstrom structure of human ferrochelatase, the terminal enzyme of heme biosynthesis. *Nat. Struct. Biol.* **2001**, *8*, 156−160.
(10) Schubot, F. D.; Chen, C. J.; Rose, J. P.; Dailey, T. A.; Dailey, H. A.; Wang, B. C. Crystal structure of the transcription factor sc-mtTFB offers insights into mitochondrial transcription. *Protein Sci.* **2001**, *10*, 1980−1988.
(11) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr.* **1998**, *D54*, 905−921.

(12) Terwilliger, T. C. Multiwavelength anomalous diffraction phasing of macromolecular structures: Analysis of MAD data as single isomorphous replacement with anomalous scattering data using the MADMRG program. *Methods Enzymol. 276*, 530−537.
(13) Lamzin, V. S.; Perrakis, A. Current state of automated crystallographic data analysis. *Nat. Struct. Biol.* **2000**, *7*, 978−981.
(14) Montelione, G. T.; Zheng, D. Y.; Huang, Y. P. J.; Gunsalus, K. C.; Szyperski, T. Protein NMR spectroscopy in structural genomics. *Nat. Struct. Biol.* **2000**, *7*, 982−985.
(15) Kwong, P. D.; Wyatt, R.; Desjardins, E.; Robinson, J.; Culp, J. S.; Hellmig, B. D.; Sweet, R. W.; Sodroski, J.; Hendrickson, W. A. Probability analysis of variational crystallization and its application to gp120, the exterior envelope glycoprotein of type 1 human immunodeficiency virus (HIV-1). *J. Biol. Chem.* **1999**, *274*, 4115−4123.
(16) Mori, S.; Abeygunawardana, C.; vanZijl, P. C. M.; Berg, J. M. Water exchange filter with improved sensitivity (WEX II) to study solvent-exchangeable protons. Application to the consensus zinc finger peptide CP-I. *J. Magn. Reson., Ser. B* **1996**, *110*, 96−101.
(17) Prestegard, J. H.; Al-Hashimi, H. M.; Tolman, J. R. NMR structures of biomolecules using field oriented media and residual dipolar couplings. *Q. Rev. Biophys.* **2000**, *33*, 371−424.
(18) Tian, F.; Valafar, H.; Prestegard, J. H. A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. *J. Am. Chem. Soc.* **2001**, *123*, 11791−11796.
(19) Fischer, D.; Elofsson, A.; Rychlewski, L.; Pazos, F.; Valencia, A.; Rost, B.; Ortiz, A. R.; Dunbrack, R. L. CAFASP2: The second critical assessment of fully automated structure prediction methods. *Proteins: Struct. Funct. Genet.* **2001**, 171−183.
(20) Rohl, C. A.; Baker, D. De novo determination of protein backbone structure from residual dipolar couplings using rosetta. *J. Am. Chem. Soc.* **2002**, *124*, 2723−2729.
(21) Valafar, H.; Prestegard, J. H. Rapid classification to a protein fold family using a statistical analysis of dipolar couplings. *Bioinformatics*, in press.
(22) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J. H.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI−BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(23) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(24) Word, J. M.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalis, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. Visualizing and quantifying molecular goodness-of-fit: Small- probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **1999**, 285, 1711−1733.

AR0101382