J Struct Funct Genomics (2009) 10:165–179
DOI 10.1007/s10969-008-9057-4

The Center for Eukaryotic Structural Genomics

John L. Markley · David J. Aceti · Craig A. Bingman · Brian G. Fox · Ronnie O. Frederick · Shin-ichi Makino · Karl W. Nichols · George N. Phillips Jr. · John G. Primm · Sarata C. Sahu · Frank C. Vojtik · Brian F. Volkman · Russell L. Wrobel · Zsolt Zolnai

Received: 18 September 2008 / Accepted: 12 December 2008 / Published online: 8 January 2009
© The Author(s) 2009. This article is published with open access at Springerlink.com

Abstract The Center for Eukaryotic Structural Genomics (CESG) is a "specialized" or "technology development" center supported by the Protein Structure Initiative (PSI). CESG's mission is to develop improved methods for the high-throughput solution of structures from eukaryotic proteins, with a very strong weighting toward human proteins of biomedical relevance. During the first three years of PSI-2, CESG selected targets representing 601 proteins from *Homo sapiens*, 33 from mouse, 10 from rat, 139 from *Galdieria sulphuraria*, 35 from *Arabidopsis thaliana*, 96 from *Cyanidioschyzon merolae*, 80 from *Plasmodium falciparum*, 24 from yeast, and about 25 from other eukaryotes. Notably, 30% of all structures of human proteins solved by the PSI Centers were determined at CESG. Whereas eukaryotic proteins generally are considered to be much more challenging targets than prokaryotic proteins, the technology now in place at CESG yields success rates that are comparable to those of the large production centers that work primarily on prokaryotic proteins. We describe here the technological innovations that underlie CESG's platforms for bioinformatics and laboratory information management, target selection, protein production, and structure determination by X-ray crystallography or NMR spectroscopy.

Keywords CESG · LIMS · PSI Materials Repository · Protein Structure Initiative (PSI) · PepcDB · Plasmid design · Protein production · Protein structure determination · TEV protease

Abbreviations

CESG	Center for Eukaryotic Structural Genomics
FTE	Full time equivalent
HSQC	Heteronuclear single quantum correlation
LIMS	Laboratory information management system
OMIM	Online Mendelian Inheritance in Man Database
PSI	Protein Structure Initiative
RDBMS	Relational database management system
WG	Workgroup

Introduction

The Protein Structure Initiative (PSI), which is funded by the National Institute for General Medical Sciences, supports a network of four production centers, ten "specialized" or technology development centers, two modeling centers, the PSI Knowledgebase, and the PSI Materials Repository. As a means of disseminating information about the program, various centers were invited to submit accounts abstracted from their 2007–2008 Annual Report to the PSI. This account describes the Center for Eukaryotic Structural Genomics (CESG). CESG focuses on the development of technology for improving success rates and lowering costs of structure determinations of human proteins related to disease or cell differentiation and proteins from families

J. L. Markley (✉) · D. J. Aceti · C. A. Bingman · B. G. Fox · R. O. Frederick · S.-i. Makino · K. W. Nichols · G. N. Phillips Jr. · J. G. Primm · S. C. Sahu · F. C. Vojtik · R. L. Wrobel · Z. Zolnai
Center for Eukaryotic Structural Genomics, Biochemistry Department, University of Wisconsin-Madison, 433 Babcock Drive, Madison, WI 53706, USA
e-mail: markley@nmrfam.wisc.edu

B. F. Volkman
Center for Eukaryotic Structural Genomics, Biochemistry Department, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226, USA

represented only in eukaryotes. We seek to expand the range of targets amenable to structure determination to membrane proteins and proteins with N- or C-terminal membrane anchors.

In choosing new targets, CESG takes three factors into account: (1) the likelihood that its structure will improve our understanding of sequence-structure relationships (sequence of the target less than 30% identical to any in a known structure), (2) the likelihood that a structure will advance our understanding of a human disease, metabolic pathway, or genetic disorder, and (3) the likelihood that the target will be produced as a folded protein amenable to structure determination. The highest ranked targets meet all three of these criteria. CESG also invites the nomination of targets by members of the scientific community. Those that are approved must be acceptable under criterion (3) and be favorable under one or both of criteria (1) and (2). In addition, CESG gives preference to targets that may lead to functional characterization of the system studied by the outside collaborator.

CESG justifies its exclusive focus on proteins from humans and other eukaryotes as follows. Eukaryotes contain a large number of sequence-structure targets (both full-length proteins and domains) that are not represented in prokaryotes. It is clear that the PSI will need to solve structures of eukaryotic proteins in order to achieve its goal of making “the three-dimensional atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences” [1]. Moreover, we foresee no real substitutes for human proteins when it comes to detailed investigations of diseases or studies aimed at understanding complex processes, such as the differentiation of human stem cells. Technological developments made in the field of eukaryotic protein production will have wide applicability in many branches of biomedical research, including antibody production, screening for molecular interactions, and drug design. Human gene products represent a grand challenge, because of their high levels of gene splicing and the huge multiplicity of protein–protein interactions known to occur. Improved methods for preparing proteins from the human and related genomes will open up structure-function investigations of these targets. The presence of clones and protocols for preparing these proteins in the PSI Materials Repository and Knowledgebase will catalyze continued studies by the biomedical community.

Eukaryotic proteins (particularly those from humans and higher vertebrates) are difficult targets for structural genomics because of a number of factors. (1) The gene models for eukaryotic proteins are poorly developed. (2) Eukaryotic proteins contain a large number of introns and are subject to alternative splicing patterns. (3) Eukaryotic proteins frequently require chaperones for proper folding. (4) Eukaryotic proteins contain considerably more regions of intrinsic disorder, and a large fraction of them ($\sim$60%) are fully natively disordered. (5) Because of difficulties in producing and solubilizing them, few structures of recombinant eukaryotic membrane proteins have been determined. CESG focuses on developing technology to overcome these difficulties. The approaches CESG has been using toward technology development and the center’s major accomplishments are summarized in Table 1. CESG endeavors to make its technology available to the scientific community. Efforts along these lines are summarized in Table 2.

Organization of CESG

The majority of CESG’s co-investigators and employees are located in the Biochemistry Department at the University of Wisconsin-Madison. This concentration of personnel has facilitated communication and resource sharing, although in the past CESG placed strains on the space and administrative infrastructure of the Department. This year the majority of CESG staff was relocated to newly remodeled laboratory space, and a new on-line purchase ordering system was instituted to streamline administrative processes. The only CESG consortium partner is at the Medical College of Wisconsin (Milwaukee, WI); members of this group participate in CESG meetings by video conferencing and visit Madison to attend important meetings.

The CESG Executive Committee (the PI, co-PIs, Bioinformatics/Crystallography Team Leader, and Project Manager) coordinates the activities of the teams, provides scientific direction, and sets long-term goals and strategies of the project. This group meets weekly in Madison to discuss agenda items, share news of recent conferences, review outside requests, authorize the creation of work groups, and surface any new concerns. Brian Volkman attends from Milwaukee via videoconferencing. Directions and initiatives from this committee are communicated to the functional teams by the investigator at the team meetings. Overall project progress and goals are shared at All-Hands meetings which are held on a quarterly basis.

CESG is organized into eight functional teams, each focused on fundamental aspects of the project. A PI or co-Investigator is responsible for the overall operation of each team; however, within each team, a PhD level scientist or an experienced administrator (Team Leader) is responsible for the day-to-day operation of the team and assists with long-range planning. CESG tracks expenses by functional teams: administration, cloning, small-scale expression testing, large-scale *E. coli* production and purification, quality assurance, cell-free protein production, X-ray

Table 1 Technology developed at the Center for Eukaryotic Structural Genomics

Laboratory management and bioinformatics
Sesame LIMS [2]
Target tracking [3]
Impact of disorder on success of determining structures [4]
Impact of low-complexity sequence on producing proteins and determining structures [5]
Efficient target scoring protocol (JCSG has adopted a similar protocol)
CESG is developing the best database of information on natively disordered eukaryotic proteins (from NMR screening) [unpublished]
Cloning, plasmid development, strain, and media development
FlexiVector cloning [6]
Vectors developed [6–10]
Autoinduction medium refinement [8]
Autoinduction medium for Se-Met labeling [11]
Autoinduction medium for ^{15}N and ^{13}C labeling [9]
Application of the technology to the efficient production of TEV protease [12]
Technology for expression and solubility screening and protein production
Control protein workgroups developed to test and verify protocols [to be published]
High-throughput cell-free screening on GenDecoder^{TM} 1000 at ∼ 2 μg scale [13–16]
Maxwell-16 for cell-based screening and purification at ∼1 mg scale [10] and Protemist^{TM} DT-II for cell-free screening and purification at ∼1 mg scale [to be published]
Comparisons of cell-free and cell-based protein production: original comparison [17]; new comparisons underway
AKTA-based semi-automated affinity and gel filtration purification at 1–100 mg scale [18]
Technology for X-ray crystallography
High-throughput crystal screening technology, including CrystalFarm Pro [to be published]
Ensemble refinements of crystal structures [19]
Automatic Crystallographic Map Interpreter (ACMI) [20]
Technology for NMR spectroscopy
HIFI-NMR fast 3D triple-resonance NMR data collection [21]
HIFI-C fast collection of coupling data for structural constraints [22]
PISTACHIO (automated assignments) [23]
LACS (automated validation) [24]
Improved analysis of NMR chemical shifts [25, 26]
PECAN (automated secondary structure determination) [27]
PINE (automated backbone and side chain assignments, secondary structure and validation) [28]
Stereo array isotope labeling (SAIL) approach for NMR structures of larger proteins [29]

crystallography, and NMR spectroscopy. Multiple sections may be in one team (cloning, small-scale expression testing, large-scale protein production, and cell-free sections make up the Protein Production Research Team). This has enabled the center to quantify supplies and labor costs for activities associated with each section.

Progress in the development of new methods, technology, approaches, and ideas for protein production and structure determination

Platform for technology assessment

CESG has developed a multi-threaded platform (Fig. 1) that covers all steps from target selection to structure determination and data deposition. The platform supports single-step cloning leading to multiple vector possibilities, two complementary methods for producing protein (*E. coli* cell-based and wheat germ cell-free), and two complementary methods for structure determination (X-ray crystallography and NMR spectroscopy). CESG uses this platform as a test-bed to evaluate the performance of new technology as it is applied to challenging targets: soluble proteins, membrane proteins, and proteins containing membrane anchors or signal sequences from the genomes of humans and higher vertebrates. The platform is interfaced to the Sesame laboratory information management system (LIMS), which collects and organizes information about targets and what is done with them (Table 4). Throughout PSI-1 (2001–2005) and PSI-2 (2005–), CESG has developed the capability of capturing of fine-grained

Table 2 Outreach activities (in PSI-2) at the Center for Eukaryotic Structural Genomics

Outside target requests: 293 submitted, 242 approved, 7 structures determined, 33 in progress
Material Transfer Agreements: 32 initiated
Collaborations on functional studies
Collaborative publications
Cell-free workshops
NMR workshops
OMIM human proteins (See Table 3 for success rates with all eukaryotic targets)
Sesame developed at CESG is the only LIMS to be used by multiple PSI centers: used by Structural Genomics of Pathogenic Protozoa Consortium (SGPP) in PSI-1 and the New York Center on Membrane Protein Structure (NYCOMPS) in PSI-2. Sesame is also used by a number of “R01”-type laboratories
PINE (1100 jobs submitted; 15% local, 65% non-UW inside USA, 20% outside USA)
NMR approach to larger proteins: SAIL structure
Depositions to PepcDB and lead in developing accompanying data
Lead role in developing protocols and data definitions with PSI Materials Repository
1945 specific clones and 8 expression plasmids transferred to the PSI Materials Repository
Training of 37 undergraduate students and 6 high school students in molecular biology and protein chemistry

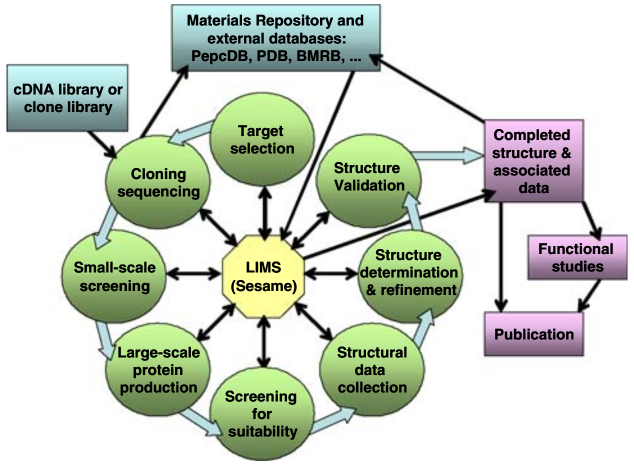
Materials Repository and external databases: PepcDB, PDB, BMRB, ...
cDNA library or clone library
Target selection
Cloning sequencing
Structure Validation
Completed structure & associated data
Small-scale screening
LIMS (Sesame)
Structure determination & refinement
Functional studies
Publication
Large-scale protein production
Screening for suitability
Structural data collection

Fig. 1 CESG’s structure determination platform interfaced with the Sesame laboratory information management system

information about individual pipeline steps into a single relational database with its Sesame LIMS. This has generated a huge database of information about the performance of eukaryotic ORFs through the various steps of our structural proteomics pipelines. The center uses this information to identify the most critical steps for improvement and to evaluate strategies for improving their efficiency.

As a technology development center in PSI-2, CESG’s approach has been to leverage its evolving protocols and strengths in protein production by cell-based and cell-free methods and structure determination by X-ray crystallographic and NMR spectroscopic methods to increase the success rates and lower the costs of determining structures of eukaryotic proteins, particularly human proteins and proteins from higher vertebrates. CESG is unique as a PSI specialized center in its ability to test new technologies at all stages of the structure determination process, including improved bioinformatics hypotheses for target selection, experimental testing of the hypotheses, extensive capture of experimental results into the Sesame LIMS system, and automated deposition of experimental results in the PSI Knowledge Base and PDB using specialized bioinformatics software. One of the lessons CESG learned from its participation in PSI-1 is that cell-based (*E. coli*) [8, 11] and cell-free (wheat germ extract) [13–16] protein production platforms offer complementary coverage in that targets that fail with one approach may succeed with the other [17]. Similarly, X-ray crystallography and NMR were found to offer complementary approaches to structure determination. Another lesson from PSI-1 was the importance of developing fast and inexpensive screens for determining the expression and solubility properties of targets that accurately predict success in subsequent scale-up [10, 13]. Much of the research in PSI-2 has focused on perfecting a robust platform that enables CESG to clone once and to carry out inexpensive small-scale trials to determine protein production, tag cleavage (if relevant), and solubility in both the cell-free and cell-based modalities.

CESG has continued to automate its X-ray crystallographic and NMR spectroscopic pipelines. With installation of a Mosquito system, CESG now can use smaller quantities of protein for crystallization trials. CESG has implemented small-scale hanging drop screening using the Mosquito for identifying suitable solution conditions for NMR spectroscopy and now uses 3 mm NMR sample tubes to reduce protein sample requirements for ^{1}H–^{15}N HSQC screening.

The robustness of CESG’s platform is evident from statistics for PSI efforts with eukaryotic proteins obtained

Table 3 Results from TargetDB, March 6, 2008 (includes PSI-1 and PSI-2)

		Selected	Work Stopped	Cloned	Expressed	Purified	X-ray	NMR	In PDB
PSI efforts with all eukaryotic proteins	All PSI	59869	14008	36350	18018	3779	267	59	313
	% Success	100%	23%	*61%*	*30%*	*6%*	< 1%	< 1%	1%
	CESG	8201	3106	8006	4254	1110	70	36	102
	% Success	100%	38%	*98%*	*52%*	*14%*	1%	< 1%	1%
CESG as a percentage of PSI effort		14%	22%	22%	24%	29%	26%	61%	*33%*
PSI efforts with human proteins only	All PSI	9310	3496	3922	2318	782	49	18	69
	% Success	100%	38%	*42%*	*25%*	*8%*	1%	1%	1%
	CESG	2099	620	2075	1300	325	13	7	20
	% Success	100%	30%	*99%*	*62%*	*15%*	1%	1%	1%
CESG as a percentage of PSI effort		23%	18%	53%	56%	42%	27%	39%	*29%*

from the NIH Knowledgebase TargetDB (Table 3). At each stage of the generic pipeline defined by the TargetDB, CESG's efforts account for a significant fraction of the total PSI effort on eukaryotes, and CESG contributed 29% of the structures of human proteins deposited in the PDB.

Laboratory information management and bioinformatics

Sesame, CESG's laboratory information management system [2], consists of a series of web-based distributed Java applications designed to organize data generated by projects in structural genomics, structural biology, and shared laboratory resources. Sesame allows collaborators on a given project to enter, process, view, and extract relevant data, regardless of location, so long as web access is available. Sesame is a multi-tier system, with data residing in a relational database. As its associated relational database management system (RDBMS), Sesame supports Oracle 8.1.7+, Microsoft SQL Server 2005, and PostgreSQL 8+ (and advanced open-source RDBMS). Sesame is by far the most complete and best-tested, publicly available LIMS. Full details concerning the Sesame project can be found at http://www.sesame.wisc.edu. Sesame currently is being used by a second PSI-2 center (the New York Consortium on Membrane Protein Structure) and by a number of other laboratories around the world.

Sesame has undergone steady development to keep up with evolving technology and the needs of the project. Modifications have been made to enable the capture of data in formats compatible with the requirements of PepcDB and the Materials Repository. The present system can handle truncated ORFs, multiple chains, and protein–ligand complexes. Data entry into Sesame has been streamlined through the introduction of tablet computers into the laboratory. Sesame allows users to import into Sesame both crystallization images and scores from the CrystalFarm database. A new "metal assay view" organizes metal analyses. Sesame now has the capability of creating "genealogy" traces that track records back to a given ORF. A new feature allows URLs to be attached to Sesame records; these provide links to bioinformatics sites and to the PDB and BMRB depositions for the target. Improved data mining tools enable a wide variety of searches and the creation of specialized reports.

A Sesame application called "Jar", which is accessible from both the CESG and Sesame websites, manages the growing number of structure requests from the community to CESG. Persons requesting structures are asked to enter information that enables CESG to quickly review the requests to decide which ones meet the guidelines for acceptance. Jar supports communications with requestors to let them know if their target has been accepted and to inform them about its progress and eventual fate: solved structure or work stopped as the result of a specified failure.

Target selection

Targets cloned by CESG in PSI-2 (1044 in total through March, 2008) included human and other mammalian proteins that have high medical relevance and other proteins expressed during differentiation of human stem cells. Human targets (601, 58%) represent the majority of all work now in progress at CESG. Work in other organisms is undertaken to test specific experimental hypothesis on ways to improve the preparation and structure determination of eukaryotic proteins. CESG routinely scans TargetDB for possible conflicts with other structural genomics centers and actively avoids duplication of effort. In the past, CESG used OMIM (Online Mendelian Inheritance in Man Database) as the major indicator for biomedical relevance. This approach was fruitful, in that publications describing structures chosen in this way had a higher impact than those chosen as fold-space targets. In several cases, it was possible to use the structure to rationalize the consequences of naturally occurring human mutations with associated phenotypes.

However, use of OMIM as the sole indicator of biomedical relevance limits the search to the 12,000 human protein sequences annotated in OMIM and ignores other types of evidence.

In order to expand the reach of CESG's biomedical relevance selection criteria, we recently explored the use of a new metric, which we call the "HEAT index". The HEAT index measures how "hot" the target is from a medically relevant standpoint. It equals the sum of: (1) the number of literature articles associated with the human protein sequence in NIH's Protein Information Resource (PIR) (ref_count); (2) the number of gene ontology annotations associated with the PIR sequence (GO_count) multiplied by 2; and (3) the number of OMIM annotations associated with the sequence (OMIM_count) multiplied by 10. Thus:

$$\text{HEAT index} = \text{Ref\_count} + 2 * \text{GO\_count} + 10 * \text{OMIM\_count} \qquad (1)$$

CESG groups newly selected targets into "Workgroups" (WG), which usually contain 96 targets but may contain fewer. All of the workgroups launched by CESG in PSI-2 (through March, 2008) are listed in (Table 4). "Selection Workgroups" are used to investigate project hypotheses, including best sources for homologs to human medically relevant proteins. "Control Workgroups" are used to test new innovations in protein production methods. "Outside Request" workgroups contain targets nominated by scientists from outside CESG. Human stem cell workgroups are collaborative efforts to evaluate novel targets, often not available in public clone collections, such as MGC, that are being identified in differentiating stem cells by our collaborator Dr. James Thomson.

One of the problems with implementing new technology is making sure that apparent gains in one step do not have deleterious effects on subsequent steps. For example, CESG found that a change made to lower costs of cloning, which changed the N-terminal cloning artifact on the purified protein, led to decreased rates of successful crystallization. To avoid such costly problems in the future, CESG instituted a panel of test proteins, called the "Control Workgroup". CESG's Control Workgroup consists of 24 targets derived from previous CESG efforts. It includes genes and proteins whose behavior is known at each step, including PCR amplification and cloning, small-scale and large-scale expression in *E. coli* cells, expression in cell-free translation, protein purification, proteolysis from fusion tags, crystallization, and HSQC spectra. Through use of the Sesame LIMS, all previous trials of these known targets can be compared with results from trials with emerging technologies. Some of the proteins in the control workgroup were chosen because they have an easily recognized chromophore or an easily assayed enzyme activity, which facilitates development of function-based screening protocols. The control workgroup provides a useful platform for validation of the routine operation of various pipeline procedures and provides a well-understood set of genes and protein targets for training new researchers. Study of these genes and proteins provides a baseline for judging the efficacy of proposed changes at any step of the research pipeline.

Table 4 Groups of targets ("Workgroups") launched by CESG in PSI-2

Workgroup designator (GE.)	Descriptive name of the workgroup
100	Outside Requests WG2: Protein from collaborators; PSI-1 follow-ups
1455	R&D WG1: Crystallization–; Ligand binding studies
1588	Selection WG1: MGC Vertebrate Clones Sequence-Structure
1589	Selection WG2: MGC Vertebrate Clones Sequence-Structure
1590	Selection WG3: MGC Vertebrate Clones Sequence-Structure
1746	Selection WG4: Galdieria Sequence-Structure
1757	Outside Requests WG3: Raines Onconase mutant; co-crystalization targets
1789	Outside Requests WG1: cDNA provided by collaborators
1811	Selection WG5: OMIM Medical Relevance Under 23 kDa
1829	Selection WG6: OMIM Medical Relevance 22–33 kDa
1855	Selection WG7: *C. merolai* Sequence-Structure
1866	Human Stem Cell WG1: Thomson/Blommel
1974	Human Stem Cell WG2: Thomson/Junying
1995	Human Stem Cell WG3: Thomson/Bradfield predicted cDNA
2147	Selection WG8: OMIM Medical Relevance; N-term signal peptide removal
2182	Outside Requests WG4: Human Rieske protein
2195	Selection WG9: Human sequence-structure less than 27 kD
2196	Selection WG10: Human sequence-structure 27–37 kDa
2350	Outside Requests WG5: PRP24 N1234 + RNA co-crystallization
2372	Control WG1: Target Selection Master
2387	Control WG3: Gateway pVP16
2394	Control WG3: pEU-His-Flexi
2398	Control WG2: Flexi Vector pVP68K
2421	Selection WG11: Galdi partners to GE.2422 Human OMIM targets
2422	Selection WG12: Human OMIM partners to GE.2421 Galdi homologs
2453	Unmodelable domains from yeast

In recognition that certain classes of human proteins may ultimately be unsuitable for structural determination, CESG postulated that thermophilic eukaryotic genomes may provide homologs with properties better suited to structure determination. A pilot project (GE.1746) investigating a few targets chosen from two eukaryotic thermophiles (“red algae”) yielded significantly higher success rates (>10%) for X-ray structure determination than achieved with human or other vertebrate targets (~2%). On the basis of these results, CESG designed a study consisting of 24 human targets of biomedical relevance (GE.2422) and 24 homologues of these targets from the thermophile *Galdieria sulphuraria* (GE.2421). All of these targets are being screened for protein production on the cell-free and cell-based platforms. Smaller proteins (>24 kDa) are initially examined as NMR targets, and if they fail in the NMR platform, they are considered for crystallization trials. Larger proteins are produced with Se-Met labeling and sent to the X-ray platform. Although this work accounts for a small portion of the CESG research effort (139 targets, 13% effort in PSI-2), it has the potential to yield important new information. If the structure of the human protein in the pair is solved, this will provide direct insights. If the structure of only the *Galdieria* homolog is solved, it will enable a model to be built of the human protein. If both structures are solved, their comparison may indicate reasons for thermostability. Even knowledge about ways to produce thermostable analogs of human proteins of biomedical relevance could be of eventual interest.

Target cloning and vector development

Vector design

In PSI-1, CESG perfected separate cloning and vector production methods for its cell-free and cell-based pathways. Although successful, the downside was the added expense of cloning and sequencing twice. CESG’s experience with eukaryotic genomes demonstrated that sequence verification of targets is an essential quality assurance step. CESG has accumulated a great depth of experience with the complications that arise from working with eukaryotic genomes, including mismatch with annotated gene models, differences in splicing, use of alternative start codons, and other biological complications. In PSI-2, CESG partnered with Promega in evaluating the Flexi®Vector cloning approach, which is less expensive than the Gateway™ approach adopted in PSI-1 and which allows a single cloning step to support both cell-based and cell-free protein expression. Because sequencing is time consuming, having a single cloning step that can support high fidelity transfer of a sequence-verified gene into many different types of expression vectors is a useful property of an integrated expression-testing platform.

The importance of performing tests of new technology over the complete set of steps comprising a structure determination pipeline is emphasized by CESG’s work with the Flexi®Vector and Gateway™ systems. Project results showed that Gateway™ was an efficient cloning system, but inclusion of the attB recombination site in the cell-free vector was inhibitory to translation. Substitution of the Flexi®Vector restriction sites allowed similar efficiency in cloning and transfer of genes to different expression plasmids. Importantly, the Flexi®Vector restriction sites were not inhibitory to the cell-free translation. With this knowledge, CESG first tested the least expensive design for primers with Flexi®Vector, which also had the consequence of leaving an N-terminal AIA amino acid sequence on the purified protein. Testing at the small-scale expression, large-scale cell growth, and protein purification stages showed that the N-terminal AIA sequence had no deleterious effect on process efficiency. However, later results indicated that the AIA tag might interfere with crystallization. Similar studies showed that inclusion of Roger Tsien’s Flash Tag motif [30] in the linker region as a possible detection scheme did not noticeably affect expression or solubility of fusion protein targets, but had a strong tendency to yield aggregated protein during aerobic purification on Ni-IMAC.

In response to these experimental evaluations obtained from realistic pipeline operations, CESG designed a new linker region and developed an integrated cloning procedure that gives the same target protein from all project vectors currently being evaluated. This new cloning approach was fully evaluated by the control workgroup approach described above, and as a consequence of the positive results, the new pVP68K plasmid has been chosen as the new standard for both cell-free and cell-based protein production platforms. Tests have shown that the PIPE cloning approach developed by the Joint Center for Structural Genomics [31] can be used for initial cloning into CESG’s small pEU-Flexi plasmid used for cell-free protein production, which contains a cleavable His-tag. As an efficient alternative, we now carry out PCR cloning into the small pEU vector, which is then sequenced. This now enables economical cell-free screening of each and every new clone for protein production and solubility. Following cleavage of the cell-free target, the sequence is identical to that produced in the *E. coli* platform following cleavage of the His-MBP-tag. Thus the solubility results of proteins produced by the cell-free and cell-based platforms are directly comparable, and differences can be related to protein folding problems rather than intrinsic solubility of the construct.

The modular design of the CESG vectors makes it easy to switch out various components. The Flexi®Vector system allows for rapid and high fidelity transfer of a cloned and sequence-verified target gene to any desired expression context. Vectors that have improved performance during auto-induction have been obtained by manipulation of the promoter for the LacI repressor protein. Bacterial expression vectors were developed that allow in vivo cleavage of MBP fusion proteins to liberate a His8-tagged target protein that can subsequently be processed with TEV protease. This vector is hypothesized to facilitate screening for solubility and automated purification. A vector that makes a fusion of the target protein with the folding chaperone protein trigger factor has been made. This vector is hypothesized to be of use expressing proteins that require additional chaperone support for folding. Each of these vectors is now being tested with targets in control workgroups (Table 4), and the results from these studies will be published.

Cell-based protein production

Refinement of the auto-induction approach

The auto-induction approach was developed initially by William Studier [32] at another PSI center. CESG adopted this approach and optimized it for the preparation of Se-Met labeled proteins [9] or stable-isotope labeled proteins [9]. Since then, CESG has used auto-induction routinely in its cell-based pipeline. In the course of work spanning nearly five years, we encountered unexplained discrepancies between yields at the initial screening and production stages, problems that had vexed attempts of other PSI centers to use this method. As a consequence, we launched a major research effort to understand this system.

The auto-induction method of protein expression in *E. coli* is based on diauxic growth resulting from dynamic function of *lac* operon regulatory elements (*lacO* and *lacI*) in mixtures of glucose, glycerol, and lactose. Our results showed that successful execution of auto-induction is strongly dependent on the plasmid promoter and repressor construction, on the oxygenation state of the culture, and on the composition of the auto-induction medium. Thus, hosts expressing high levels of LacI during aerobic growth exhibit reduced ability to effectively complete the auto-induction process. Manipulation of the promoter to give decreased expression of LacI altered the preference for lactose consumption in a manner that led to increased protein expression and partially relieved the sensitivity of the auto-induction process to the oxygenation state of the culture. We employed factorial design methods to optimize the chemically defined growth medium used in the production of two model proteins, *Photinus* luciferase and enhanced green fluorescent protein, either as unlabeled or selenomethionine-labeled proteins. The optimizations included studies of protein production from T7 and T7-*lacI* promoter plasmids and from T5 phage promoter plasmids expressing two levels of LacI. From the analysis of over 500 independent expression results, we identified combinations of optimized expression media and expression plasmids that yielded greater than 1000 µg of purified protein per ml of expression culture. Figure 2 shows one plane of a 3D space defined by changes in glucose, glycerol, and lactose concentrations. The media compositions identified by these optimizations, which were markedly different than those published by Studier [32], gave yields of recombinant GFP as high as 1.5 mg per milliliter of culture fluid.

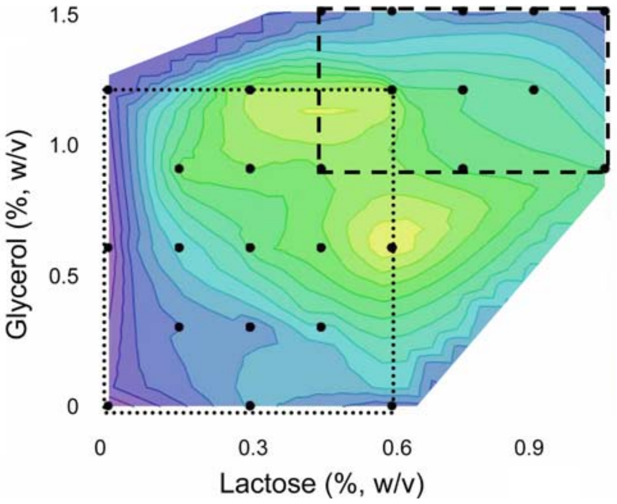

Fig. 2 Topographical map representing protein production results from a factorially evolved auto-induction medium. Lower production levels are indicated by *blue hues* and higher levels by *yellow hues*. Experimental design points are shown as *black circles*

Use of the expression vector modified to attenuate LacI repressor levels and the optimized auto-induction medium improved the correlation between the scores from small-scale screening (aerobic) and large-scale production (oxygen limited) from ~50% observed with the prior methods to ~80%. The newly optimized conditions also yield higher volumetric cell mass. The factorial array of media conditions includes compositions that yield auto-induction in early log phase growth, mid-log growth or late-log growth. These results may explain literature reports that auto-induction can be either beneficial or deleterious to the expression of individual proteins. The availability of the factorial auto-induction medium array allows these anecdotal reports to be investigated in a more systematic manner.

CESG has begun to partner with non-PSI groups to apply these approaches to their own problems. Some topics of interest include peptide display methods using custom-designed MBP-thioredoxin fusion proteins, mammalian Rieske proteins, full-length human cytochrome b5 reductase, and soluble human cytochrome b5 reductase. In each case, application of CESG methods has led to robust high-level expression and, for the proteins containing cofactors, high-level incorporation. These latter results continue our experience with lactose-based auto-induction during expression of metalloproteins dating back to 1994 [33].

Small-scale expression screening and protein production

In 2007, we published a study showing how vector design and medium evolution could be linked to a rapid 1 ml-scale purification using the Promega Maxwell IMAC-based purification robot [10]. The approach was used to prepare a Se-Met-labeled protein sample sufficient for an X-ray structure determination. In addition, the method has been used to prepare labeled proteins for NMR structural studies in a one-day growth and purification cycle. Each protein was produced from a 50 ml culture, and the total cost of each ^{15}N-labeled protein was ~$50 ($28 for purification and $22 for media). A single Maxwell robot (~$20 K) is capable of performing up to 128 purifications per day, making this an ideally matched technology for meso-scale (0.2–2 mg) protein production. This technology also is appropriate for individual investigators and has been disseminated already to laboratories at UC San Diego and the Medical College of Wisconsin. As part of this work, we are continuing to implement a microfluidic, automated protein analysis using a Caliper LabChip 90 to replace SDS-PAGE methods.

Protein purification

The protein purification capabilities developed in PSI-1 focused on protocol definition and developing simple automation for the ÄKTA Purifier platform. Over the period of PSI-2, we have developed the capacity to perform the routine purification of up to 12 protein targets per week, including extensive data capture at each step of the purification process. We use a combination of Ni^{2+}-IMAC and preparative gel filtration to prepare all samples from cell-based expression. With the improved correlation in cell growth phase described above, we are now beginning to examine correlations between small-scale predictions and the final protein yield, purity, and other target quality indicators. These results will be used to further refine our understanding of the cell-based production pipeline. Drop-frozen samples of purified protein are made available for quality assurance testing (SDS-PAGE, UV–vis, mass spectra). Upon satisfactory completion of the quality assurance testing, the protein samples are made available for transfer to the structure determination groups.

The purification team has begun research on the use of isoelectric focusing as a specialized polishing purification step. For proteins containing cofactors or metal ions that might be removed by Ni-IMAC chromatography, CESG has begun to use maltose binding protein affinity as the first purification step. This approach has been used successfully to purify proteins containing iron-sulfur centers, including soluble Rieske-type ferredoxins from human and mouse brain. The presence of a soluble Rieske ferredoxin in a mammal was not known before this work. The structure showed that the protein is most closely related to the electron carriers from bacterial multi-component oxygenase systems. The function of the mammalian protein is not yet known. The purification group has also begun studies on liposome floating as a specialized purification technique for membrane proteins produced in both *E. coli* and cell-free translation. We have shown how this approach can be used to prepare highly purified liposomes containing full-length functional human cytochrome b5.

New efforts on membrane proteins

CESG's *E. coli* expression vectors have been shown to express full-length human cytochrome b5 with the C-terminal membrane anchor as a soluble fusion protein. We developed methods to purify the fusion protein in the absence of detergents and to deliver the full-length functional protein in situ to liposomes upon treatment with TEV protease. Figure 3a summarizes this useful approach. The work has been published [34] and is included in a patent application. Based on this success, CESG will further investigate the use of this approach for the preparation of other N- and C-terminal anchored membrane proteins in a systematic manner.

The liposome preparation method has also been studied using cell-free translation (Fig. 3b). Proteins currently selected for study include bacteriorhodopsin (expression control); stearoyl-CoA desaturases from human, mouse, and *Mycobacterium tuberculosis*; sigma-1 receptor (human, guinea pig, and rat); and laminin-α, β and γ from human stem cells. The Wisconsin Alumni Research Foundation (WARF) has filed patents on production of stearoyl-CoA desaturase by this approach.[1] We have found that cell-free translation in the presence of liposomes gives

[1] B. G. Fox, P. Sobrado and Y. Chang. 2006. Cell-free expression systems for mammalian and mycobacterial desaturases with utility for drug screening. WARF Case No: P06127US. Wisconsin Alumni Research Foundation. September 1, 2006; B. G. Fox and Y. Chang. 2006. Novel mycobacterium tuberculosis protein. WARF Case No: P06129US. Wisconsin Alumni Research Foundation. September 1, 2006; B. G. Fox and Y. Chang. 2007. Inhibition of *Mycobacteria smegmatis* C-terminal tail specific protease (msTsp) and its utility in enhancement of heterologous expression WARF Case No: 08214US. Wisconsin Alumni Research Foundation).

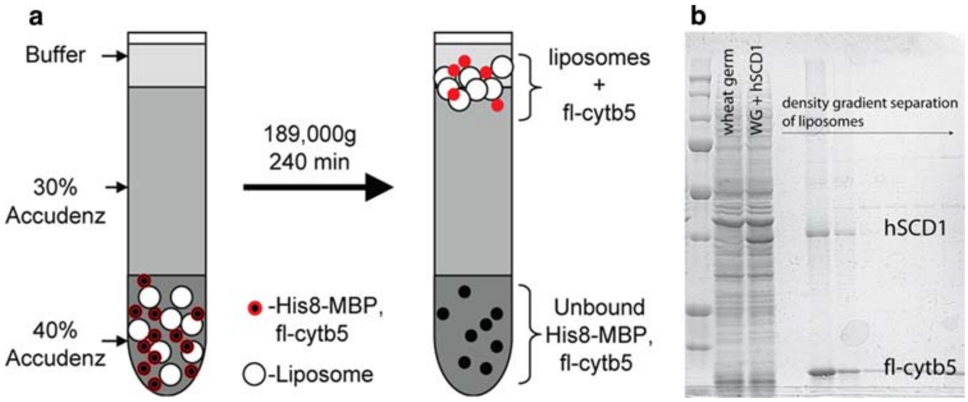

Fig. 3 **a** Density gradient capture of cytochrome b5 into liposomes after in situ proteolysis of fusion protein purified from *E. coli* or after wheat germ cell-free translation. **b** SDS PAGE analysis of the density gradient separation of integral membrane human stearoyl-CoA desaturase in the presence of exogenously added human cytochrome b5

high level expression and near complete transfer of the expressed proteins to liposomes. We are collaborating with Dr. Lloyd Smith (University of Wisconsin Chemistry Department) to develop efficient protease digest and mass spectral approaches to map the orientation of expressed membrane proteins in the liposome bilayer. In the case of sigma-1 receptor, ligand binding studies undertaken in collaboration with Dr. A. Ruoho (University of Wisconsin Pharmacology Department) confirm that a fraction of the expressed and lipid incorporated protein is capable of binding ligands, demonstrating a functional state. In the case of stearoyl-CoA desaturase, the enzyme is a three-protein complex. We have been able to use cell-free translation express human stearoyl-CoA desaturase and simultaneously incorporate exogenously added human cyt b5 (Fig. 3b). Methods have been developed to introduce the diiron center into the desaturase, incorporate heme into cytb5, and the N-terminal anchored cytochrome b5 reductase has been added to the liposome-bound complex to reconstitute the 3-protein integral membrane complex. Catalytic assays using [^{14}C-]-stearoyl-CoA show the in vitro complex is catalytically active. Moreover, the expressed protein purified by the liposome floating method is ~80% pure by capillary electrophoresis. We are now working on methods to exchange the liposome for detergents. These efforts are designed to build experience and analytical procedures suitable for handling a larger set of membrane proteins proposed by CESG as structural genomics targets.

Cell-free protein production

Wheat germ cell-free protein production has proven to be an efficient and economical method for screening targets for expression and solubility and for making labeled proteins for NMR structural studies. CESG's cell-free platform is highly automated. It requires the efforts of only two FTEs for screening, protein production, and purification, as compared to five FTEs for the cell-based platform. The overall cell-free process also is much faster at every step than cell-based.

Previous studies comparing *E. coli* cell-based and wheat germ cell-free approaches to preparing labeled proteins for NMR studies showed that nearly twice as many folded and soluble proteins resulted from the cell-free approach [17]. Wheat germ cell-free has become a robust platform at CESG for the production of labeled proteins for NMR structure determination [16]. In the past year, CESG has investigated the production of membrane proteins from wheat germ extracts in the presence of detergent micelles or liposomes. A large proportion of the membrane proteins tried expressed well and were solubilized under these conditions. These preliminary results suggest that this may provide a pathway to producing proteins for structure determination by NMR spectroscopy or X-ray crystallography.

CESG makes use of a Cell-Free Sciences GenDecoder1000^{TM} robotic system in automating the small-scale screening of constructs for protein production and solubility. This unit makes it possible to carry out as many as 384 *small-scale* (25-μl) screening reactions per 48-h run. The average yield is 2–5 μg/well, sufficient for determining expression levels and solubility by comparing PAGE from total and soluble protein or for determining function through an enzymatic activity or ligand binding assay. Current average supplies costs, including wheat germ extract are: $4/target for expression-solubility testing and analysis.

Targets that express well in small scale are next screened at intermediate scale on the Protemist^{TM} DT-II robotic system (CellFree Sciences). The Protemist^{TM} DT-II carries out fully automated transcription, translation, and batch method affinity purification. Translation is carried out in bilayer mode, with wheat germ extract and mRNA in the bottom layer and amino acids and energy source in the top layer. Six 6-ml samples can be produced in 35 h, including purification, with a yield of 0.1–0.3 mg protein per sample (0.6–1.8 mg from all six samples). The cost is ~$80/sample for unlabeled protein or $90/sample for a ^{15}N-labeled sample. The only additional steps required for producing an

NMR sample are buffer exchange and concentration. The yield is sufficient for buffer screening to determine proper solution conditions for NMR and for initial screening by NMR spectroscopy to determine if the protein is folded and monodisperse.

Large-scale cell-free protein production is carried out on one of CESG's two robotic units. The Protemist^{TM} 10 robotic system is capable of carrying out eight 4-ml translation reactions per 24 h run; this system requires preparation of the mRNA off-line. The Protemist^{TM} 100 robotic system supports eight 4-ml transcription and translation reactions per 48-h. Typical yields for the Protemist runs are 0.3–0.5 mg purified protein per ml of reaction. The cell-free purification protocols generally require less time and effort than the corresponding cell-based ones, owing to smaller initial volumes and higher initial purity. Using the latest in GE HIS-TRAP purification technology, IMAC purification of His-tagged proteins requires 40 min of processing time and results in protein samples that are 75–85% pure. Gel filtration adds an additional 3 h and can increase the purity to >95% for proteins <15 kD and to 90% for proteins <20 kD. GST purification results in >95% purity regardless of size; however, the minimal time to process the sample is greater than 10 h. These systems are used to make ^{15}N- and ^{13}C,^{15}N-labeled samples for NMR structure determination or Se-Met samples for X-ray structure determination. The average yield is 0.2–0.4 mg of purified protein/ml of reaction. Current average supplies costs per 4-ml reaction (yielding 0.8–1.2 mg protein), including wheat germ extract, labeled amino acids, and purification, is $390 for [^{15}N]-protein, $370 for Se-Met–protein and $470 for [^{13}C,^{15}N]-protein.

CESG has been successful in using the Protemist^{TM} DT-II cell-free robotic system to translate membrane proteins in the presence of detergent micelles. One of these proteins has been exchanged into a detergent that is suitable for NMR studies and has been shown to yield a promising NMR spectrum (Fig. 4). NMR structural studies on this target are underway.

NMR data collection and analysis

In collaboration with staff members of the National Magnetic Resonance Facility at Madison (NMRFAM), CESG has developed novel automated approaches to protein NMR data collection and analysis. An adaptive and interactive engine for identifying peaks in 3D triple-resonance spectra in a probabilistic manner is based on ideas of reduced-dimensionality and tilted-plane data collection. The method, called "High-Resolution Iterative Frequency Identification" (HIFI)-NMR [21], eliminates the spectral reconstruction step and concentrates on finding the best model for the peak positions. Optimal 2D spectra are chosen on the fly. In so doing, the algorithm can in 2 h create automatically a statistically annotated peak list that would take over 20 h to produce by conventional means. The HIFI software is tasked with the job of identifying the chemical shifts of all the peaks that would appear in a 3D spectrum. After each new 2D spectrum is acquired at a different angle, the software reviews all data and updates its model. It also judges when all possible peaks have been captured. The current HIFI-NMR software can collect data from the set of NMR experiments needed to determine the solution structure of a small protein and turn these into peak lists for the next step in the analysis. The peak lists generated by HIFI-NMR along with the sequence of the protein are input to a software package called PISTACHIO (Probabilistic Identification of Spin sysTems and their Assignments including Coil-Helix Inference as Output) [23]. The output of PISTACHIO is chemical-shift assignments represented as a series of minimum energy configurations ranked according to their probable correctness, with multiple attributions in ambiguous cases. With typical data sets, initial PISTACHIO runs yield reliable backbone assignments for more than 90% of the amino acid residues and reliable side chain assignments with greater than 75% completeness. The assigned data are then run through the LACS [24] (Linear Analysis of Chemical Shifts) software, which corrects possible referencing problems and identifies assignment outliers. The software package PECAN [27] (Protein Energetic Conformational ANalysis) then carries out probabilistic secondary structure determination.

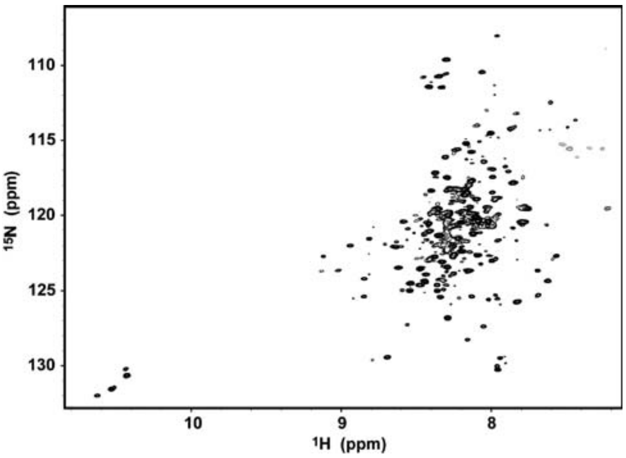

Fig. 4 ^{1}H–^{15}N TROSY HSQC spectrum (600 MHz) of a ^{15}N-labeled membrane protein produced in the wheat germ cell-free system containing detergent (Brij35) and subsequently exchanged into 0.5% Fos-choline-12 (lipid-like zwitterionic detergent). The protein contains four Trp residues, whose signals are clearly resolved in the lower left corner of the spectrum. The sample temperature was 25°C. The spectrum is the average of 256 transients

The separate data analysis steps (PISTACHIO, LACS, and PECAN) have now been fully integrated into a pipeline, called PINE-NMR. PINE handles input data from a wide variety of NMR experiments, including HCCH TOCSY (important for side chain assignments). PINE is freely available from a web server (http://miranda.nmrfam.wisc.edu/PINE/). PINE has been used by NMR spectroscopists worldwide for over 1,000 assignment runs. Current efforts are focused on linking HIFI-NMR and PINE so that full data analysis follows directly from fast data collection.

A HIFI-NMR approach to the collection of reduced dipolar coupling (RDC) data from a partially oriented protein, called HIFI-C [22], can accelerate the collection of these data by a factor of 3–5. The increase in speed is important, because many proteins are unstable in oriented media needed for their partial orientation and begin to precipitate over time. Another advantage of this approach is that the splittings are determined multiple times (from data at different tilted angles) so that error bars can be associated with the RDC values.

Technology development for X-ray crystallography

Crystallization screening

The database and automation-centric approach of CESG to identification of crystal leads and optimization of diffraction quality crystals has continued to mature. In PSI-1, and continuing through the second year of PSI-2, CESG assembled all components of a high-throughput pipeline, and optimized them individually and as an integrated platform. Recently, a Mosquito plate setup robot has been added, which lowers the required crystallization volumes from 1 μl to 100 nl. The Mosquito produces droplets that are well suited for automatic image classification, reduces the sample requirements for screening by a factor of 10, allows us to screen in depth with less sample, and expands the range of screened conditions, to include additive screens.

Collaborative work between the Sesame development team and the crystallography group continues to build on the robust platform already in place. Most notably, a transport mechanism has been developed for moving scores and images from the CrystalFarm databases to Sesame. This constitutes a critical step toward CESG's goal of uploading all crystallomics data and images to PepcDB.

Significant research effort was expended toward analyzing data from our 48-condition additive screen and expanding it to 96 conditions. The current formulation of our additive screen can be described as a "compacted" version of our original cofactor heavy additive screen into a few cocktails, and incorporation of additional reagents known to facilitate protein crystallization. The current additive screen also includes a number of complex mixtures of natural products. Additionally, reagents intended to modify protein side-chains have been included. Wide-scale implementation of the new additive screen is possible primarily because of the acquisition of the Mosquito robot. Support for capturing the data from additive screening was completed in year 1 of PSI-2.

Once monodisperse proteins are produced, the bottleneck for solving crystal structures is the crystallization process. This is generally viewed as a combinatorial or incomplete factorial design process, with thousands of experiments required to obtain suitable crystals for a structure determination. Robotics have been integrated into many steps of the process, including the management of solutions for crystallization, setting up of trays for crystal screening, plate handling for time series image capture, crystal lead identification, and optimization of crystal conditions. This latter step often benefits from the experience of crystallographers, and there is no standard algorithm. To aid in the process of converting conceptual principles to practical experiments in protein crystal optimization, we have developed a tool called Crystal Farm Pro, a Java application which queries the database of the Crystal Farm imaging system, presents screening data to the crystallographer, and generates solution conditions for optimization experiments based on the contents of original screening conditions. The design of the software also includes hooks to implement liquid handler control scripts to robotically generate the solutions for optimization plates from stock solutions. Progress in the last year includes its deployment at several third party sites, better network connectivity, new data mining and statistical analyses of crystallization outcomes, and various improvements in error handling.

CESG's crystallization imaging needs continue to be well served by a pair of 400-position CrystalFarm imaging robots. Our collaborative agreement with Bruker has let to important advances in our CrystalPro tool. Significantly, a built-in statistical test now provides pair-wise comparisons of crystallization outcomes, as will be needed to facilitate analysis of data from the control workgroup, which has progressed to crystallization screening. Another new feature supports deep data-mining of crystallization outcomes, across all instances of a given screen over the entire CrystalFarm database. The presentation of data on crystallization outcomes has been improved by the provision of detailed statistical reports on constituents associated with positive crystallization outcomes for a given target. These reports, which facilitate the optimization of leads, will be especially useful in analyzing data from our new additive screen.

During year 1 of PSI-2, our project provided a database of scored images to Bruker and Discovery Partners to facilitate the automatic classification of crystallization images. The challenge of automatic crystallization image classification has now been taken up by a collaboration between Phillips, Amos Ron (Computer Science, UW-Madison) and Robert Nowak (ECE UW-Madison.) A previously developed baseline algorithm has been adapted to function with images produced with current best practice (100–200 nanoliter droplets set with the Mosquito robot and acquired with the CrystalFarm imaging system). Efforts to improve the robustness of the methodology and sharpen the classification are ongoing.

In-house production of reagents

As an important part of efforts to control costs, CESG has used its own protein production platform to produce protein reagents for project use. The refined auto-induction and vector design approaches were first applied to tobacco etch virus NIa proteinase (TEV protease) production. TEV protease is an important tool for the removal of fusion tags from recombinant proteins. From this work, CESG corrected numerous incorrect statements in the literature about the relative stability of various TEV protease constructs, and obtained expression of TEV protease at high levels and with high solubility by using auto-induction medium at 37°C. In combination with the expression work, an automated two-step purification protocol was developed that yielded His-tagged TEV protease with >99% purity, high catalytic activity, and purified yields of ~400 mg/l of expression culture (~15 mg pure TEV protease per g of *E. coli* cell paste). Methods for producing glutathione S-transferase tagged TEV with similar yields (~12 mg of pure protease fusion per g of *E. coli* cell paste) were also published. Since publication of this work [11], numerous laboratories from all over the world have requested these vectors, including several NIH PSI centers. These same approaches have also been used to obtain purified, highly active human rhinovirus 3C protease, and the first requests for this plasmid have begun to arrive.

The Flexi®Vector cloning system requires the use of a high concentration of T4 DNA ligase. Thus, T4 DNA ligase has become a major cost. As an approach to lowering this cost, we tested the possibility of making the ligase in house, as we do with TEV protease. An *E. coli* expression plasmid was obtained for a His6 tagged T4 DNA ligase using a new project vector. Preliminary experiments showed that it expressed well in our auto-induction medium and yielded a protein of high purity after only one IMAC purification step. The isolated protein is as active as the commercially available high concentration ligase in the standard assay used by Promega. If the in-house prepared ligase proves to works well in our Flexi®Vector cloning protocol, we will adopt this approach to lowering costs.

Discussion

CESG is a leader in bringing structural genomics approaches to bear on human and other eukaryotic proteins. The understanding gained will assist in successfully leveraging PSI-derived structural genomics methods into more complicated, but highly relevant, biological topics such as alternative splicing patterns, temporal patterns of gene expression in stem cell differentiation, tumor cell biogenesis, and many other biological phenomena that are unique to eukaryotes.

CESG worked with the PSI Materials Repository (PSI-MR) and the University of Wisconsin to develop acceptable material transfer agreements for this important aspect of sharing the results of the PSI. CESG and the PSI-MR also developed required protocols defining the format of the electronic documents to be transferred between the depositor and the PSI-MR prior to shipment of plasmids, establishing the string of check-offs that must occur before plasmids are shipped, and developing definitions and descriptors for experimental results associated with each plasmid deposition. All depositors to the PSI-MR will use these check-offs. As a result of these efforts, CESG was the first PSI center to transfer materials to PSI-MR.

The Sesame LIMS system was modified to include actions to track PSI-MR sample selection, all phases of preparation, and availability of the target to the general public though the PSI-MR. After the electronic processes were deemed acceptable, CESG deposited 96 target clones into the PSI-MR as a pilot study for the physical transfer process. This pilot project was successfully concluded when we were able to re-order some of our plasmids from the PSI-MR and receive them back at CESG. CESG has currently transferred 1945 targets to the PSI-MR, and we anticipate clearing our backlog of clones by May, 2009. Further refinements in information exchange between PSI-MR and CESG are ongoing.

CESG is the now the first PSI center to provide target-specific experimental results such as g of cell paste used in the purification, mg yield of protein, and other experimental observables captured in the Sesame LIMS system. At this time, no other PSI center is providing this level of detail, which will be essential for the biological community to make broader use of the NIH Knowledge Base. CESG has twenty-two technology dissemination reports available through PSI-KB at http://cci.lbl.gov/kb-tech/. All CESG peer reviewed publications (total of 130) have been registered with the PSI-KB.

Acknowledgments The authors thank all members of the Center for Eukaryotic Structural Genomics, past and present, for their contributions to this project. This work was supported by the NIH, Protein Structure Initiative grant U54 GM074901. The National Magnetic Resonance Facility at Madison, which collaborated in NMR technology development and provided data collection support is funded by NIH grants P41 RR02301 and P41 GM66326.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Norvell JC, Machalek AZ (2000) Structural genomics programs at the US National Institute of General Medical Sciences. Nat Struct Biol 7(1):931. doi:10.1038/80694
2. Zolnai Z, Lee PT, Li J, Chapman MR, Newman CS, Phillips GN Jr, Rayment I, Ulrich EL, Volkman BF, Markley JL (2003) Project management system for structural and functional proteomics: Sesame. J Struct Funct Genomics 4:11–23. doi:10.1023/A:1024684404761
3. Pan X, Wesenberg GE, Markley JL, Fox BG, Phillips GN Jr, Bingman CA (2007) A graphical approach to tracking and reporting target status in structural genomics. J Struct Funct Genomics 8:209–216. doi:10.1007/s10969-007-9037-0
4. Oldfield CJ, Ulrich EL, Cheng Y, Dunker AK, Markley JL (2005) Addressing the intrinsic disorder bottleneck in structural proteomics. Proteins 59:444–453. doi:10.1002/prot.20446
5. Bannen RM, Bingman CA, Phillips GN Jr (2007) Effect of low-complexity regions on protein structure determination. J Struct Funct Genomics 8:217–226. doi:10.1007/s10969-008-9039-6
6. Blommel PG, Martin PA, Wrobel RL, Steffen E, Fox BG (2006) High efficiency single step production of expression plasmids from cDNA clones using the Flexi Vector cloning system. Protein Expr Purif 47:562–570. doi:10.1016/j.pep.2005.11.007
7. Thao S, Zhao Q, Kimball T, Steffen E, Blommel PG, Riters M, Newman CS, Fox BG, Wrobel RL (2004) Results from high-throughput DNA cloning of Arabidopsis thaliana target genes using site-specific recombination. J Struct Funct Genomics 5:267–276. doi:10.1007/s10969-004-7148-4
8. Blommel PG, Becker KJ, Duvnjak P, Fox BG (2007) Enhanced bacterial protein expression during auto-induction obtained by alteration of lac repressor dosage and medium composition. Biotechnol Prog 23:585–598. doi:10.1021/bp070011x
9. Tyler RC, Sreenath HK, Singh S, Aceti DJ, Bingman CA, Markley JL, Fox BG (2005) Auto-induction medium for the production of [U-^{15}N]- and [U-^{13}C, U-^{15}N]-labeled proteins for NMR screening and structure determination. Protein Expr Purif 40:268–278. doi:10.1016/j.pep.2004.12.024
10. Frederick RO, Bergeman L, Blommel PG, Bailey LJ, McCoy JG, Song J, Meske L, Bingman CA, Riters M, Dillon NA, Kunert J, Yoon JW, Lim A, Cassidy M, Bunge J, Aceti DJ, Primm JG, Markley JL, Phillips GN Jr, Fox BG (2007) Small-scale, semi-automated purification of eukaryotic proteins for structure determination. J Struct Funct Genomics 8:153–166. doi:10.1007/s10969-007-9032-5
11. Sreenath HK, Bingman CA, Buchan BW, Seder KD, Burns BT, Geetha HV, Jeon WB, Vojtik FC, Aceti DJ, Frederick RO, Phillips GN Jr, Fox BG (2005) Protocols for production of selenomethionine-labeled proteins in 2-L polyethylene terephthalate bottles using auto-induction medium. Protein Expr Purif 40:256–267. doi:10.1016/j.pep.2004.12.022
12. Blommel PG, Fox BG (2007) A combined approach to improving large-scale production of tobacco etch virus protease. Protein Expr Purif 55:53–68. doi:10.1016/j.pep.2007.04.013
13. Vinarov DA, Lytle BL, Peterson FC, Tyler EM, Volkman BF, Markley JL (2004) Cell-free protein production and labeling protocol for NMR-based structural proteomics. Nat Methods 1:149–153. doi:10.1038/nmeth716
14. Vinarov DA, Markley JL (2005) High-throughput automated platform for nuclear magnetic resonance-based structural proteomics. Expert Rev Proteomics 2:49–55. doi:10.1586/14789450.2.1.49
15. Vinarov DA, Loushin Newman CL, Tyler EM, Markley JL (2006) Protein production using the wheat germ cell-free expression system. Curr Protoc Protein Sci 1–18
16. Vinarov DA, Loushin Newman CL, Markley JL (2006) Wheat germ cell-free platform for eukaryotic protein production. FEBS J 273:4160–4169. doi:10.1111/j.1742-4658.2006.05434.x
17. Tyler RC, Aceti DJ, Bingman CA, Cornilescu CC, Fox BG, Frederick RO, Jeon WB, Lee MS, Newman CS, Peterson FC, Phillips GN Jr, Shahan MN, Singh S, Song J, Sreenath HK, Tyler EM, Ulrich EL, Vinarov DA, Vojtik FC, Volkman BF, Wrobel RL, Zhao Q, Markley JL (2005) Comparison of cell-based and cell-free protocols for producing target proteins from the Arabidopsis thaliana genome for structural studies. Proteins 59:633–643. doi:10.1002/prot.20436
18. Jeon WB, Aceti DJ, Bingman CA, Vojtik FC, Olson AC, Ellefson JM, McCombs JE, Sreenath HK, Blommel PG, Seder KD, Burns BT, Geetha HV, Harms AC, Sabat G, Sussman MR, Fox BG, Phillips GN Jr (2005) High-throughput purification and quality assurance of Arabidopsis thaliana proteins for eukaryotic structural genomics. J Struct Funct Genomics 6:143–147. doi:10.1007/s10969-005-1908-7
19. Levin EJ, Kondrashov DA, Wesenberg GE, Phillips GN Jr (2007) Ensemble refinement of protein crystal structures: validation and application. Structure 15:1040–1052. doi:10.1016/j.str.2007.06.019
20. DiMaio F, Kondrashov DA, Bitto E, Soni A, Bingman CA, Phillips GN Jr, Shavlik JW (2007) Creating protein models from electron-density maps using particle-filtering methods. Bioinformatics 23:2851–2858. doi:10.1093/bioinformatics/btm480
21. Eghbalnia HR, Bahrami A, Tonelli M, Hallenga K, Markley JL (2005) High-resolution iterative frequency identification for NMR as a general strategy for multidimensional data collection. J Am Chem Soc 127:12528–12536. doi:10.1021/ja052120i
22. Cornilescu G, Bahrami A, Tonelli M, Markley JL, Eghbalnia HR (2007) HIFI-C: a robust and fast method for determining NMR couplings from adaptive 3D to 2D projections. J Biomol NMR 38:341–351. doi:10.1007/s10858-007-9173-7
23. Eghbalnia HR, Bahrami A, Wang L, Assadi A, Markley JL (2005) Probabilistic identification of spin systems and their assignments including coil-helix inference as output (PISTACHIO). J Biomol NMR 32:219–233. doi:10.1007/s10858-005-7944-6
24. Wang L, Eghbalnia HR, Bahrami A, Markley JL (2005) Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. J Biomol NMR 32:13–22. doi:10.1007/s10858-005-1717-0
25. Wang L, Eghbalnia HR, Markley JL (2006) Probabilistic approach to determining unbiased random-coil carbon-13 chemical shift values from the protein chemical shift database. J Biomol NMR 35:155–165. doi:10.1007/s10858-006-9022-0
26. Wang L, Eghbalnia HR, Markley JL (2007) Nearest-neighbor effects on backbone alpha and beta carbon chemical shifts in proteins. J Biomol NMR 39:247–257. doi:10.1007/s10858-007-9193-3
27. Eghbalnia HR, Wang L, Bahrami A, Assadi A, Markley JL (2005) Protein energetic conformational analysis from NMR

chemical shifts (PECAN) and its use in determining secondary structural elements. J Biomol NMR 32:71–81. doi:10.1007/s10858-005-5705-1

28. Bahrami A, Markley JL, Assadi A, Eghbalnia HR (2009) Probabilistic interaction network of evidence: application to key steps in the automation of protein structure determination by NMR spectroscopy. PLoS Comp Biol (in press)
29. Takeda M, Sugimori N, Torizawa T, Terauchi T, Ono AM, Yagi H, Yamaguchi Y, Kato K, Ikeya T, Jee JP, Güntert P, Aceti DJ, Markley JL, Kainosho M (2008) Structure of the putative 32 kDa myrosinase binding protein from Arabidopsis (At3g16450.1) determined by SAIL-NMR. FEBS J 275:5873–5884. doi:10.1111/j.1742-4658.2008.06717.x
30. Griffin BA, Adams SR, Jones J, Tsien RY (2000) Fluorescent labeling of recombinant proteins in living cells with FlAsH. Methods Enzymol 327:565–578. doi:10.1016/S0076-6879(00)27302-3
31. Klock HE, Koesema EJ, Knuth MW, Lesley SA (2008) Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. Proteins 71:982–994. doi:10.1002/prot.21786
32. Studier FW (2005) Protein production by auto-induction in high-density shaking cultures. Protein Expr Purif 41:207–234. doi:10.1016/j.pep.2005.01.016
33. Borggrefe T, Davis R, Erdjument-Bromage H, Tempst P, Kornberg RD (2002) A complex of the Srb8, -9, -10, and -11 transcriptional regulatory proteins from yeast. J Biol Chem 277:44202–44207. doi:10.1074/jbc.M207195200
34. Sobrado P, Goren MA, James D, Amundson CK, Fox BG (2008) A protein structure initiative approach to expression, purification, and in situ delivery of human cytochrome b5 to membrane vesicles. Protein Expr Purif 58:229–241. doi:10.1016/j.pep.2007.11.018