Center for Eukaryotic Structural Genomics

Protein Structure Initiative

Establishment of Pipeline Procedures for Eukaryotic Structural Genomics

Center for Eukaryotic Structural Genomics (CESG), University of Wisconsin-Madison, Madison, WI 53706-1549, USA, http://www.uwstructuralgenomics.org

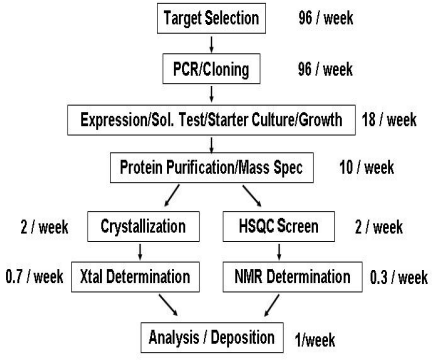
Abstract

This poster presents an overview of the Center for Eukaryotic Structural Genomics (CESG). More detailed information about parts of the project are presented in five additional posters on: "High-Throughput Protein Production for Structural Investigations from a Cell-Free Wheat Germ Expression System," "High-Throughput Production of Unlabeled, Selenomethionine- and ^{15}N-labeled Proteins in Terrific Broth and Chemically Defined Auto-Induction Media," "High-Throughput Purification of *Arabidopsis thaliana* Proteins Overexpressed in *E. coli* for Structural Genomics," "High-Throughput Methods for Producing ^{15}N- and ^{13}C; ^{15}N-Labeled *Arabidopsis* Proteins from *E. coli* Cells, for Screening for Optimal Solution Conditions, and for Investigating their Suitability for NMR Structural Analysis," and "High-Throughput Crystallomics at the Center for Eukaryotic Structural Genomics."

CESG has begun to generalize its pipeline procedures, which have been developed in the context of targets selected from the *Arabidopsis thaliana* genome, to include targets from other eukaryotic genomes. The most recent release of CESG's Sesame LIMS system is now capable of handling targets representing full-length proteins or protein fragments (domains) from multiple genomes. CESG is developing the capability of evaluating an enlarged universe of targets in terms of its tested target selection criteria. The choice of targets is influenced by information concerning the availability of the gene, either from a particular cDNA library, as supported by gene chip data or from a resource that supplies sequenced clones. The next stage of the project will be to evaluate how procedures for structural genomics of *Arabidopsis* (described below) translate to targets from other eukaryotic genomes.

Flowchart of the CESG Pipeline

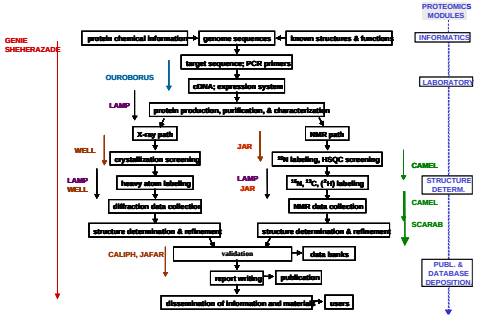
The following flowchart represents an abbreviated work flow strategy used by CESG for year 3 of the project.. Points of interest include the use of *Arabidopsis thaliana* callus tissue cell culture as the transcript source, cloning by Invitrogen's recombinational "Gateway" methods, protein production using *E. coli* expression systems, and a simplified purification system using Ni-IMAC.



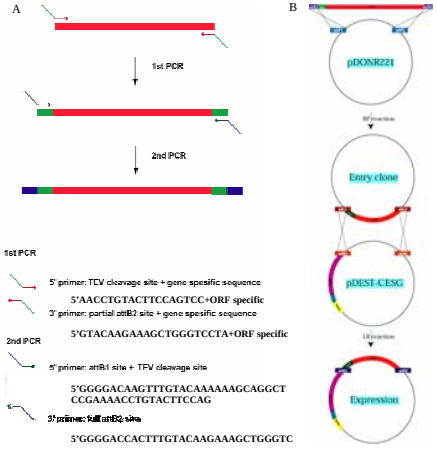
SESAME: A Laboratory Information Management System

CESG utilizes the SESAME program developed in NMRFAM, UW-Madison, to track all experimental steps. From the homepage (**http://kamba.nmrfam.wisc.edu/Sesame/**), modules can be accessed for data storage and management of target ORFs, expression vectors, protein production, NMR samples, NMR experiments, scheduling of NMR spectrometers, and crystallization screens. **Bottom**: Genie is used for select targets, generating primers and tracking progress. This page of the Genie module is used for working with an individual "workgroup" (a set of target ORFs, usually numbering 96, that are processed as a unit). Actions and results describing the progress of individual ORFs from amplification from cDNA through structure determination can be entered.

CESG's Pipeline and the Relevant Sesame (LIMS Modules That Manage Information and Processing

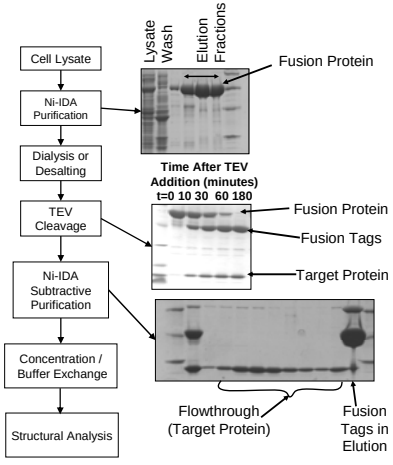


CESG uses ligation-independent cloning (Gateway system) to produce entry clones and specialized destination clones for its *E. coli-based* protein production pipeline. The destination clone fuses the target gene to a sequence coding for N-terminal (His)x-MBP followed by a TEV cleavage site. Small-scale expression testing of this construct in *E. coli* shows that ~60% of cloned *Arabidopsis* genes are expressed as a soluble fusion protein. Targets that pass this screen move to large-scale protein production. Bacterial cells are grown in plastic soda bottles with yields up to 20 g / liter of wet packed cells. Terrific broth is used as the medium for production of unlabeled protein, and a variant of the self-inducing medium developed by W. Studier is used for the production of labeled proteins (Se-Met, ^{15}N, or ^{13}C, ^{15}N). Proteins are isolated and purified in two steps by semi-automated metal affinity chromatography: first the fusion protein is isolated and subjected to TEV protease cleavage; then the uncleaved protein, His-tagged-TEV protease, and His-tagged MBP are removed. Analysis of ~1200 trials with *Arabidopsis* targets shows that the cumulative process starting with a correctly sequenced entry clone leads to pure, soluble protein with a success frequency of ~17%. Protein yields are ~1 mg to greater than 100 mg, depending on the target.



Protein Purification and Tag Removal

Our protein purification scheme relies on the Nickel binding ability of the His6-tag. The tags, including an MBP solubility tag, are cleaved from the purified fusion protein using a His6-tagged version of TEV protease. The fusion tags and the TEV are separated from the target protein through second round of Ni-IDA purification.



Status of *Arabidopsis* ORFs Targeted By CESG

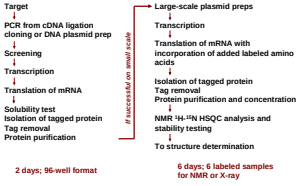
Below are progress statistics generated by the Genie module of our LIMS. Screening refers to the small-scale expression testing of the expression clones. Production refers to the large-scale growth in PET-bottles.

	NUMBER		SUCCESS RATE (Step)	
Selected Targets	1991			
PCR +	1520		76%	
Entry Clone +	1365		92%	
Destination Clone +	1194		96%	
Sequence +	1154		91%	
Screening Expression +	812	11%	70%	18%
Large-Scale Cell Growth +	639		96%	
Production Scale Expression +	588		92%	
Soluble Product +	518		82%	
Tag Cleaved +	399		77%	
Successfully Purified +	205		77%	
Crystallized	58			
Diffraction Quality Crystal	24			
Crystal Structure	10			
HSQC Data Collected	59			
HSQC --> Folded Protein	31			
NMR Structure	8			

In cooperation with Ehime University (Matsuyama, Japan) and Cell-Free Sciences (Yokohama, Japan), CESG is investigating the potential of wheat germ cell-free expression as a technology for semi-automated expression screening and protein production for structural studies by NMR and X-ray crystallography. CESG's overall success rate with the cell-free technology in going from a sequenced clone to soluble protein has been 50% for (His)6-tagged proteins (n = 146 trials) and 49% for GST-fusions including recovery of the target protein following protease cleavage (n = 102). The cell-free approach offers distinct advantages for preparing labeled proteins and currently is providing most of the ^{15}N and ^{13}C;^{15}N-labeled proteins entering the NMR structure determination pipeline.

Cell-Free Science 'GeneDecoder 1000' Installed January, 2004 at CESG - Madison

Work-Flow Diagram for Wheat Germ Cell-Free Protein Production



National Magnetic Resonance Facility at Madison, Wisconsin

CESG's NMR pipeline is adopting pre-screening of solution conditions for protein solubility as a prerequisite for ^{15}N-HSQC NMR analysis to determine the suitability of the target for solution structure analysis; about half of the ^{15}N-labeled proteins prepared as candidates for NMR structure determinations (Mr < 20,000) have passed this screen.

CESG's crystallization pipeline employs robotic screening using fixed custom screens and stochastic screens. The percentage of proteins crystallized from proteins prepared as described above is 37%. About 2/3 of these proteins eventually yield diffraction quality crystals.

The products of CESG's efforts thus far include the Sesame LIMS system, which has been deployed successfully at a second structural genomics site, written protocols describing CESG's pipeline procedures, 819 sequenced clones, 462 constructs that produce soluble protein, 239 purified proteins, 7 PDB entries with structural coordinates, and 2 BMRB entries with NMR data associated with a structure.

CESG is supported by the National Institute of General Medical Sciences through the Protein Structure Initiative (P50 GM 64598).