Center for Eukaryotic Structural Genomics:
Facility for Structure Determination of Proteins from *Arabidopsis thaliana* and Other Model Eukaryotes

Han, Byung Woo{1} Aceti, David{1} Amasino, Rick{1} Arangarasan, Raj{1} Bingman, Craig{1} Blommel, Paul{1} Buchan, Blake{1} Burch, Heather{1} Cao, John{1} Cornelescu, Claudia{1} Doreleijers, Jurgen{1} Dyer, Dave{1} Eghbalnia, Hamid{1} Fox, Brian{1} Fredrick, Ronnie{1} Geetha, Holalkere{1} Gopalakrishnan, Premkum{1} Hegeman, Adrian{1} Hruby, Dave{1} Jeon, Won Bae{1} Johnson, Ken{1} Kimball, Todd{1} Kjer, Kelly{1} Kunert, John{1} Lee, Peter{1} Ji, Jing{1}, Markley, John{1} Narayama, Ramya{1} Newman, Craig{1} Olson, Andrew{1} Phillips, George{1} Ramirez, Bryan{1} Ravoof, Nitin{1} Rayment, Ivan{1} Rosenberg, Nathan{1} Runnels, Michael{1} Seder, Kory{1} Shaw, Jeff{1} Smith, David{1} Song, Jikui{1} Sreenath, Hassan{1} Sussman, Michael{1} Thao, Sandy{1} Tyler, Ejan{1} Ulrich, Eldon{1} Vinarov, Dmitriy{1} Votjik, Frank{1} Wesenberg, Gary{1} Westler, Milo{1} Wrobel, Russell{1} Zhang, Jianhua{1} Zhao, Qin{1} Zolnai, Zsolt{1} Volkman, Brian{2} Peterson, Francis{2} Lytle, Betsy{2} Dunker, Keith{3} Oldfield, Chris{3} Linial, Michal{4} Endo, Yaeta{5} Sawasaki, Tatsuya{5} Kainosho, Masatsune{6}

{1} University of Wisconsin, Madison {2} Medical College of Wisconsin, Milwaukee {3} Molecular Kinetics, Pullman, Washington {4} Hebrew University, Jerusalem, Israel {5} Ehime University, Matsuyama, Japan {6} Tokyo Metropolitan University, Tokyo, Japan.

Abstract

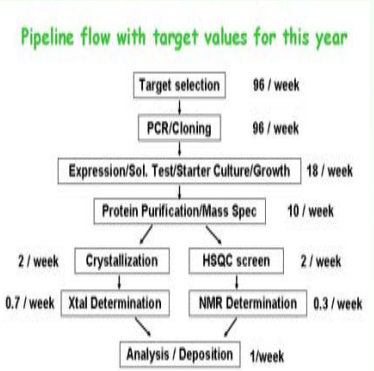
The Center for Eukaryotic Structural Genomics (CESG) was founded as a collaborative effort to develop critical technologies for determining three-dimensional structures of proteins rapidly and economically. CESG's initial focus is on the genome of the model plant *Arabidopsis thaliana*. CESG uses a laboratory information management system ('Sesame') designed to track and evaluate steps in the process leading from gene to published structure. CESG's software periodically analyzes the entire *Arabidopsis* genome to determine targets to be produced. Priority is given to targets likely to open up important regions of conformational space or to elucidate novel fold-function relationships. CESG also considers proteins of structural interested by the plant science community. Gene chips produced by maskless array DNA synthesis are being used to determine the presence of targets in cDNA pools generated by RT-PCR of RNA isolated from an Arabidopsis callus cell line. CESG's standard pipeline protocol utilizes Invitrogen's 'Gateway' plasmid construction system in 96 well plates, expression and solubility assays, large-scale *E. coli* fermentation in 2 liter disposable bottles, TEV protease-cleavable tags, semi-automated purification technology, and robotics-based crystallization. Efforts are also underway to produce protein by cell-free methods for efficient isotopic labeling for NMR structure determination. Preliminary results on the cloning, expression, solubility, and structural characterization of targets on both bacterial and cell-free systems will be presented at the meeting. Additional information can be found at http://uwstructuralgenomics.org/.

Introduction

A long-range goal in high-throughput structural biology will be to more completely map the naturally occurring diversity in protein structures. In this approach, the initial emphasis has been directed toward proteins whose sequences suggest a novel fold, proteins associated with novel functions, or proteins likely to have a known fold but a function not previously associated with that fold. In order to achieve the logistical demand of these ambitious programs, supporting methodologies that allow for specialized gene cloning, expression evaluation, and protein purification will be required. High resolution structure determination studies by NMR spectroscopy and X-ray crystallography will require the production of milligram amounts of soluble, native, and pure proteins. These efforts first become feasible by the wealth of information arising from genomic sequencing efforts, where the assumed proper identification of open reading frames has set the stage for massive scale gene cloning, followed by expression testing, and purification. The major aims of the University of Wisconsin-Madison Center for Eukaryotic Structural Genomics (CESG) are the development and critical analysis of methods for high-throughput, proteome-scale, eukaryotic protein production, characterization, and structure determination. Here we describe some of the strategies and rationale that form the basis for our high-throughput protein production efforts.

FLOWCHART OF THE CESG PIPELINE

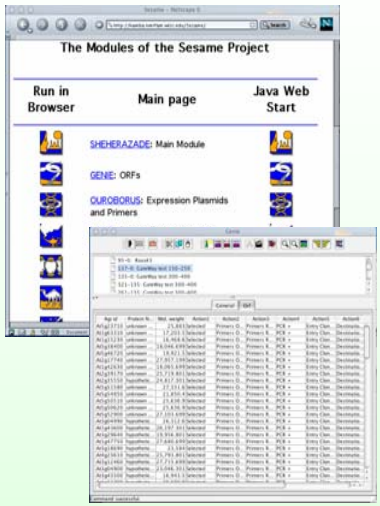
The following flowchart represents an abbreviated work flow strategy used by CESG. Points of interest include the use of *Arabidopsis thaliana* callus tissue cell culture as the transcript source, cloning by Invitrogen's recombinational "Gateway" methods, protein production using *E. coli* expression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



SESAME: A LABORATORY INFORMATION MANAGEMENT SYSTEM

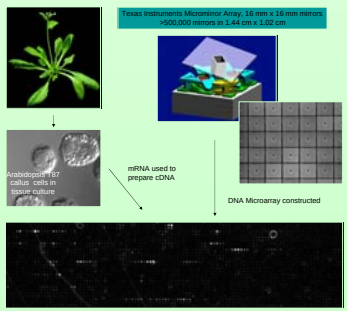
CESG utilizes the SESAME program developed in NMRFAM, UW-Madison, to track all experimental steps.

Top: From this 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.



PROFILING GENE EXPRESSION IN *ARABIDOPSIS* T87 CELLS

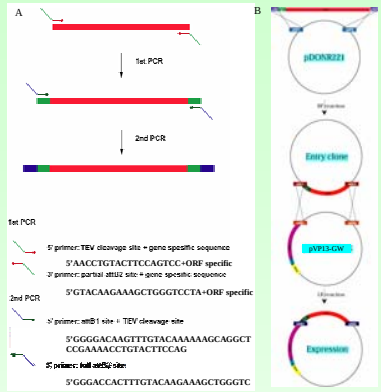
Gene chips produced by NimbleGen's maskless array DNA synthesis technology are being used to determine the presence of cDNA targets. These analyses showed that about 60-80% of all *Arabidopsis* genes may be expressed by the T87 callus cell line. These results are being integrated into GENIE to optimize workgroup generation and to guide researchers in cDNA target selection.



GATEWAY CLONING AT CESG

CESG uses Gateway technology (Invitrogen) to generate expression vectors that provide simple swapping of expression systems and protein tags. CESG has developed expression vectors that add S-tag (for detection), a His_6-tag (for purification), and MBP (maltose-binding protein, for solubilization and purification) to the N-terminal of the target protein. When required, the entire fusion is cleavable from the protein target by TEV protease. These vectors are derived from pET (T7 promoter; Novagen), pQE (T5 promoter; Qiagen), and pBAD (*AraBAD* promoter; Invitrogen) vector backbones

A. Two Step amplification of an ORF from cDNA by PCR with the addition of recombination (*att*) and TEV protease cleavage sites. B. Recombinational cloning of the PCR-generated insert to give entry and expression clones.

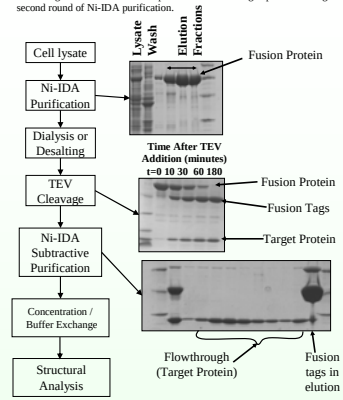


HIGH-THROUGHPUT BACTERIA GROWTH IN PET-BOTTLES

To express the heterologous proteins, *E. coli* cultures are grow in disposable polyethylene terephthalate (PET) soda-bottles in 500 ml of broth or defined media. These bottles are used to increase the culture capacity of the shaking incubators and to obviate the handling and sterilization of glassware. This system of *E. coli* growth was developed in Mark Donnelly's lab at Argonne National Laboratory (personal communication).

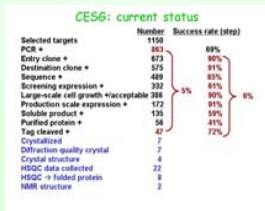
PROTEIN PURIFICATION AND TAG REMOVAL

Our protein purification scheme relies on the Nickel binding ability of the His_6-tag. The tags, including an MBP solubility tag, are cleaved from the purified fusion protein using a His_6-tagged version of TEV protease. The fusion tags and the TEV are separated from the target protein through second round of Ni-IDA purification.

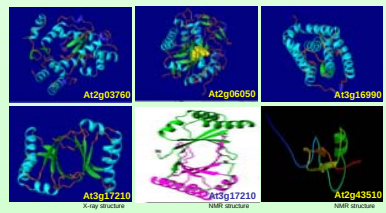


STATUS OF ARABIDOPSIS ORFs TARGETED BY CESG

Below are progress statistic revealed 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.



MOLECULAR STRUCTURES SOLVED BY CESG



CONCLUSION

CESG is using *Arabidopsis thaliana* as a model system to develop methodologies for high-throughput eukaryotic protein production and high-resolution structure determination by NMR spectroscopy and X-ray crystallography. Members of the Maize Genetics Community can submit requests for *Arabidopsis* ORFs to be solved by submitting an ORF request form found at:

http://uwstructuralgenomics.org/submission.htm.

Researchers wanting to submit Maize ORFs for structural studies can do so by providing us with a purified protein or the ORF in an Invitrogen Entry plasmid containing the same 5' and 3' sequence additions as described in this poster. Protocols for producing these entry clones are available upon request. Requested ORFs must meet our target selection criteria, the most important being that no homologous structure has been determined. Researcher must also agree to our requirement of immediate publication of the structure information. Please see our our website for additional information.

CESG is supported by the National Institute of General Medical Sciences through the Protein Structure Initiative (P50 GM64598)