Center for Eukaryotic Structural Genomics

Protein Structure Initiative

An Overview of the *Arabidopsis* Protein Production Pipeline at the Center for Eukaryotic Structural Genomics (CESG)

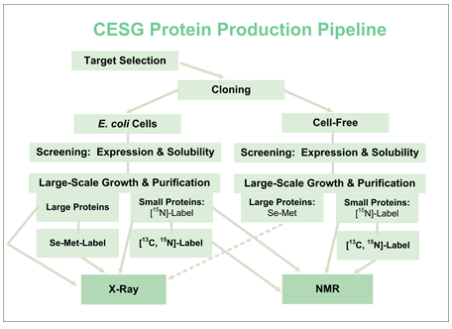
Russell L. Wrobel, Michael R. Sussman, Brian F. Volkman, Brian G. Fox, George N. Phillips, Jr., and John L. Markley

University of Wisconsin-Madison, Department of Biochemistry, 433 Babcock Drive, Madison, WI, USA 53706-1549, http://www.uwstructuralgenomics.org

Abstract

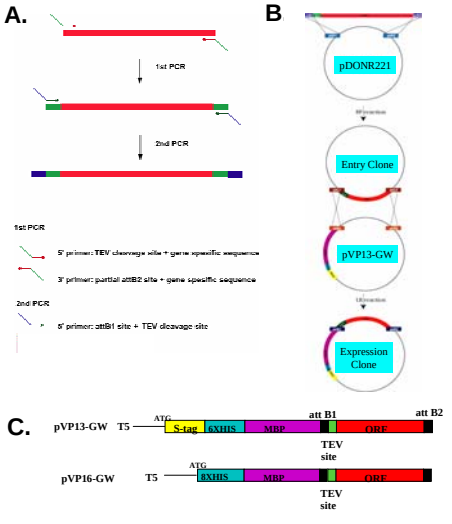
CESG was founded to develop high throughput and cost effective methods to determine the structures of eukaryotic proteins. CESG is a coordinated effort of nine centers funded by NIH and seeks to sample protein fold space to determine structure/function relationships and facilitate sequence/fold predictions. CESG initially focused on *Arabidopsis* proteins, since a high quality, thoroughly annotated genomic sequence was available, and a large percentage of the sequences are 'pioneer,' with no known clues on their function. Using semi-automated cloning methods we've cloned over 1500 *Arabidopsis* ORFs in Gateway entry vectors. These have been expressed in *E. coli* using custom-made expression vectors with amino-terminal tags to facilitate soluble expression, visualization, and purification of recombinant protein. *E. coli* strains have been developed to express these proteins in an inducing medium that also efficiently incorporates selenomethionine into the polypeptide chain to aid in X-ray diffraction analysis or ^{15}N followed by ^{15}N+^{13}C for NMR analyses. To date we have purified over 250 *Arabidopsis* targets through semi-automated methods at average yields of more than 10 mg. In a more recently developed parallel pathway, *Arabidopsis* ORFs are expressed in a wheat germ cell-free system. Currently, yields are sufficient only for NMR analyses, but we are testing chip-based crystallization technology that should enable small scale (100 µg) crystallization screening for X-ray analysis. Cell-free translation reactions are automated using screening and preparative scale robotic platforms. Up to 192 crystallization conditions are set and screened by a robot, and another robotic system optimizes protein crystallization conditions. Synchrotron data from well diffracting crystals are collected at Argonne National Laboratory. NMR data are collected at the National Magnetic Resonance Facility at Madison (NMRFAM) and the Medical College of Wisconsin (MCW).

CESG has solved the structures of 43 *Arabidopsis* proteins by NMR and X-ray crystallography. Most of these proteins were of unknown function. In several cases the structure suggested a biological or biochemical function. Please see poster Extra Abstract #3 for the *Arabidopsis* protein structures that CESG has generated and poster #650 for the resources that CESG will make available to the scientific community.



Gateway Cloning at CESG

CESG uses Gateway technology (Invitrogen) to generate expression vectors that provide simple swapping of expression systems and protein tags. **A**) Diagrammed is the two-step amplification of an ORF by PCR with the addition of recombination (*att*) and TEV protease cleavage sites. The template for the first PCR can be genomic DNA for intronless genes, bulk cDNA from reverse transcribed cellular RNA, or plasmid DNA containing cloned cDNAs. **B**) Recombinational cloning of the PCR-generated insert to give Entry and Expression clones. **C**) CESG has developed multiple Gateway compatible expression vectors. Diagramed here are the vectors we rely on most frequently. They are T5, LacI repressible promoter (pQE) based vectors. pVP13-GW has an S-tag (for visualization) and a 6XHIS-tag (for purification). pVP16-GW has an 8XHIS tag (for purification) and both have the MBP (maltose-binding protein, for solubilization) fused to the N-terminal of the target protein. When required, the entire fusion is cleavable from the protein target by TEV protease. Please see poster #650 by Craig Newman for a list of the *Arabidopsis* clones generated by CESG.

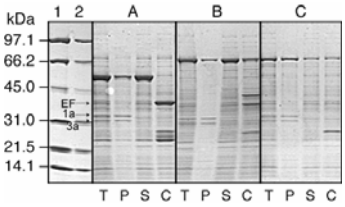


Protein Production in *E. coli*

CESG has developed an *E. coli* based protein production pipeline that incorporates the auto-induction strategy introduced by Studier for the production of SeMet-labeled proteins in *Escherichia coli* B834. Expression constructs are first screened for expression in 1 ml cultures. Expression positive constructs are then grown in 2-L polyethylene terephthalate (PET) beverage bottles. In practice, each bottle contains 500 mL of culture medium. CESG currently uses four culture bottles per protein target to meet project-defined production goals.

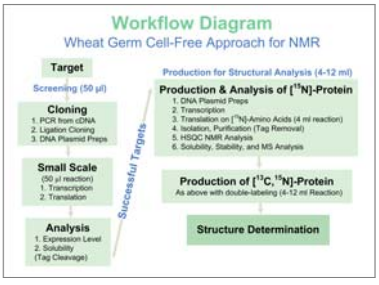
Scoring of Protein Expression

CESG uses three-tier scoring to evaluate: (1) total protein expression, (2) the level of soluble expression, and (3) the percentage of total soluble fusion protein that can be proteolyzed by tobacco etch virus (TEV) protease. Analysis of total protein expression, solubility, and proteolysis of fusion protein three example targets using denaturing gel electrophoresis. The marked lanes were used to evaluate total expression (*T*), soluble expression (*S*), insoluble or pellet (*P*), and TEV cleavage (*C*).



Cell-Free Protein Production

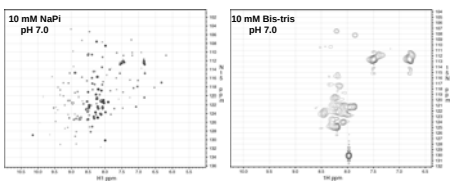
CESG has developed an automated platform for NMR-based structural proteomics that employs wheat germ extracts for cell-free production of labeled proteins. The platform utilizes a single construct with non-cleavable (N)-His_6 for all targets. The GeneDecoder 1000™ robot is used for small scale [^{15}N] labeled protein production used for screening and the Protemist™ is used for large scale [^{13}C, ^{15}N]-protein production. A construct slated for cell-free expression must be cloned using another approach, restriction/ligation cloning, since the cell-free expression is incompatible with Gateway Technology.



Gene Decoder1000™ Protemist10™

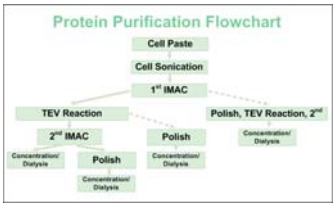
HSQC NMR Analysis

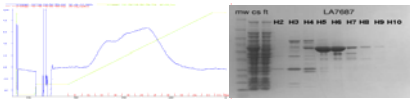
As an initial screen, HSQC NMR analysis is used to determine the folded state and stability of protein samples. The below spectra shows that the protein At1g23750, a predicted DNA binding protein, was unfolded in a Bis-tris buffer but was stably folded in phosphate buffer at the same pH.



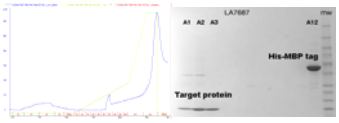
Protein Purification Pipeline

The protein purification pipeline at CESG provides highly purified proteins from A*rabidopsis* genome. Using an automated ÄKTA Purifier system, fusion proteins are initially purified by 1st IMAC capture. After cleavage of (His)n-MBP tags by TEV protease, (His)n-MBP tags are separated from target proteins by 2nd IMAC capture. Generally, the purity of most of the target proteins was greater than 90%. When the purity of target protein was less than 80%, polishing steps using either ion exchange and/or size exclusion improved the purity to greater than 90%. Polishing programs were established with an ÄKTA HPLC system





Chromatogram and SDS-polyacrylamide gel of the 1st IMAC capture of a typical His-tagged MBP fusion protein.

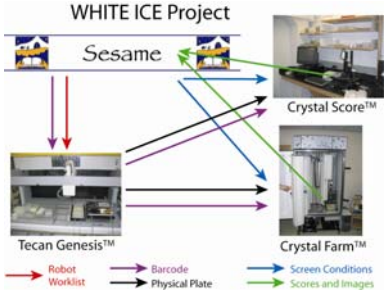


Chromatograms and SDS-polyacrylamide gels of the 2nd IMAC removal of His-MBP tags from a typical target proteins.

Crystallomics at CESG

WHITE ICE *Wisconsin High-Throughput Extensible and Integrated Crystallization Environment*

A highly integrated environment has been developed and implemented for the generation of crystals for CESG studies. Robotics and associated database tools allow for the management of crystal stock solutions, initial screening, imaging, scoring, and optimization — all coupled to the Sesame laboratory information management system. The Well module of Sesame functions as the control center for coordinating crystallization activities at CESG.



Screening, Optimization, and Salvage

CESG has adopted the Corning Crystal EX plate for initial screening. This 192-condition vapor-diffusion microtiter-format plate gives optimum performance in robotic setup and imaging. All screening and optimizations are performed using the worklists written by the Well module of Sesame, and executed on the Tecan Genesis™.

Imaging and Scoring

CESG has used two semi-automated CrystalScore™ systems. CrystalScore™ is now the ambient pipeline imaging system. CrystalFarm™ is fully automated and capable of storing up to 400 plates and imaging them on a pre-determined schedule or on demand. CrystalFarm™ is our imaging platform for 4ºC crystallization experiments.

Acknowledgments

Members and collaborators of the CESG are: **UW-Madison:** Dave Aceti, Rick Amasino, Raj Arangarasan, Arash Bahrami, Craig Bingman, Paul Blommel, Blake Buchan, Heather Burch, John Cao, Claudia Cornilescu, Gabriel Cornilescu, Jurgen Doreleijers, Dave Dyer, Hamid Eghbalnia, Ronnie Fredrick, Holalkere Geetha, Premkum Gopalakrishnan, Byung Woo Han, Adrian Hegeman, Dave Hruby, Won Bae Jeon, Ken Johnson, Todd Kimball, Kelly Kjer, John Kunert, Min Lee, Peter Lee, Jing Li, Scott Leisman, Miron Livny, Andrew Markley, Zach Miller, Ramya Narayama, Craig Newman, John Primm, Brian Ramirez, Nitin Ravoof, Ivan Rayment, Megan Riters, Michael Runnels, Kory Seder, Mark Shahan, Jeff Shaw, Shanteri Singh, David Smith, Jikui Song, Hassan Sreenath, Sandy Thao, Donna Troestler, Ejan Tyler, Robert Tyer, Eldon Ulrich, Dmitriy Vinarov, Frank Vojtik, Liya Wang, R. Kent Wenger, Gary Wesensberg, Milo Westler, Jianhua Zhang, Qin Zhao, Zsolt Zolnai; **Medical College of Wisconsin:** Betsy Lytle, Brian Volkman, Francis Peterson, **Molecular Kinetics:** Keith Dunker, Chris Oldfield; **Hebrew University (Jerusalem):** Michal Linial, Elon Portugaly, Ilone Kifer; **German National Center for Health & Environment (Munich):** Dmitrij Frishman; **Tokyo Metropolitan University:** Masatsune Kainosho, Yuko Katagiri, Nozomi Sugimori, Akira M. Ono, Tsutomu Terauchi, Takuya Torizawa; **Ehime University**: Yaeta Endo, Tatsuya Sawasaki; **CellFreeSciences, Inc. (Yokohama):** Ryo Morishita, Mihoro Saeki, Motoo Watanabe.

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