THE CENTER FOR EUKARYOTIC STRUCTURAL GENOMICS AT THE UNIVERSITY OF WISCONSIN-MADISON

Frits Abildgaard, [1]David J. Aceti, [1]Craig Bingman, [1]Paul Blommel, [1]Heather Burch, [1]John Cao, [1]Claudia Cornilescu, [1]Jurgen Doreleijers, [1]David Dyer, [1]Hamid Eghbalnia, [1]Brian G. Fox, [1]Ronnie O. Frederick, [1]Adrian Hegeman, [1]Won Bae Jeon, [1]Todd Kimball, [1]Kelly Kjer, [1]Peter T. Lee, [1]Jing Li, [2]Betsy Lytle, [2]John L. Markley, [1]Ramya K. Narayana, [1]Craig S. Newman,[1]George N. Phillips, [1]Francis Peterson, [2]Ivan Rayment, [1]Bryan Ramirez, [1]Mike J. Runnels, [1]Kory Seder, [1]David Smith, [1]Mike Sussman, [1]Sandy Thao, [1]Eldon L. Ulrich, [1]Dmitriy Vinarov, [1]Brian F. Volkman, [2]Gary Wesenberg, [1]W. Milo Westler, [1]Russell L. Wrobel, [1]Jianhua Zhang, [1]Qin (Kate) Zhao,[1]Zsolt Zolnai

[1]Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Drive, Madison, Wisconsin 53706, and [2]Department of Biochemistry, Medical College of Wisconsin, 8701 Watertown Plank Road,Milwaukee, Wisconsin 53226

ABSTRACT

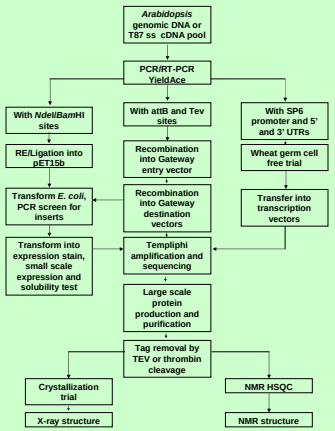
The major objectives of the University of Wisconsin Center for Eukaryotic Structural Genomics (CESG) are the development and critical analysis of methodologies for high-throughput, proteome-scale, eukaryotic protein production using *Eschericia coli* and cell-free translation, protein characterizations such as S-Tag detection, solubility, activity, and mass spectral analysis, and high resolution structure determination. The ultimate goal is to develop high-throughput structural biology methodologies to aid in the visualization of the complete diversity in naturally occurring protein structures. In this poster we summarize some of the strategies and rationale of our high-throughput protein production efforts as applied to the model photolithotropic eukaryote, *Arabidopsis thaliana*. In this system, protein targets were chosen from a prioritized list of about ~25,000 predicted *Arabidopsis* genes: we have focused on proteins whose primary sequences suggest novel folds or known folds accompanied by unusual biological functions. To meet the needs of this ambitious project, numerous supporting methodologies that allow for specialized gene cloning, gene expression evaluation, and protein purification are being developed. Finally, high-resolution structure determination of proteins is being solved using both NMR spectroscopy and X-ray crystallography methods.

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 Center for Eukaryotic Structural Genomics 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 STRATEGY

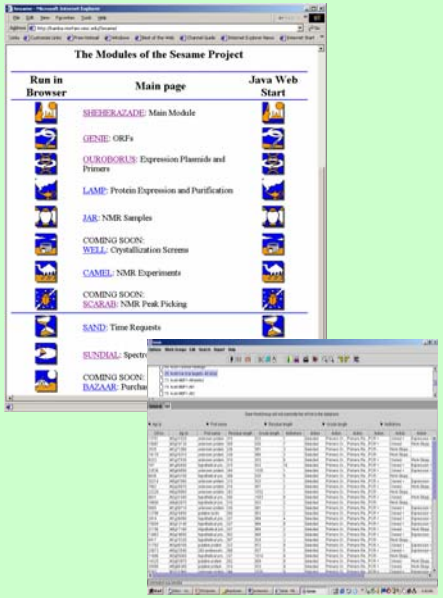
The following flowchart represents the strategy being developed at CESG. Points of interest include the use of *Arabidopsis thaliana* callus tissue cell culture as the transcript source, cloning by both restriction/ligation and recombinational methods, wheat germ cell-free protein synthesis as an alternative to *E. coli* heterologous expression, and Templiphi (Amersham Biosciences Sunnyvale, California) to generate DNA for sequencing.



SESAME: A LABORATORY INFORMATION MANAGEMENT SYSTEM

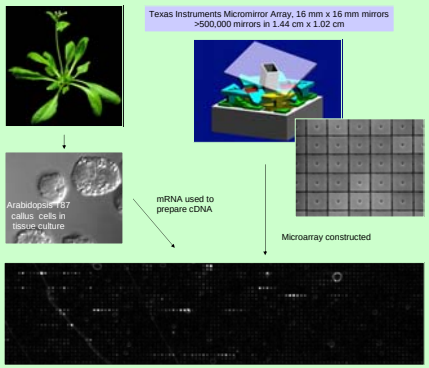
CESG utilizes the SESAME program developed in NMRFAM, UW-Madison, to track all experimental steps.

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

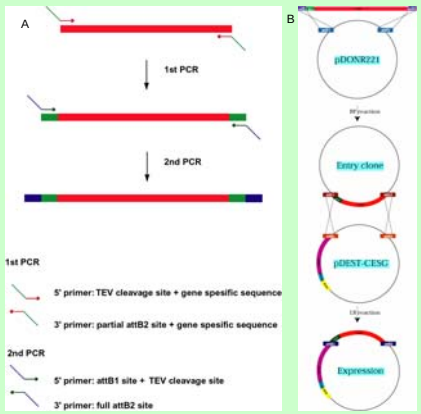
About 60-80% of all *Arabidopsis* genes may be expressed by the T87 callus cell line upon the basis of DNA microarray analyses. These results are being integrated into GENIE to optimize workgroup generation and to guide researchers in the selection of further experimental conditions.



GATEWAY CLONING AT CESG

CESG uses Gateway technology (Invitrogen) to generate expression vectors that provide simple swapping of expression contexts for the target protein.

A. 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.



EXPRESSION VECTORS DEVELOPED AT CESG

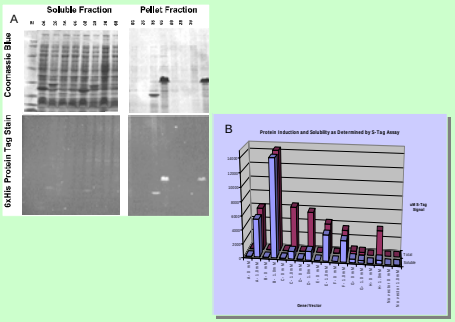
CESG has developed expression vectors that add S-tag (for detection), a His_6-tag (for purification), and MBP (maltose-binding protein, for solubilization and purification) to the N-terminal of the target protein. When required, the entire fusion is cleavable from the protein target by TEV protease. These vectors are derived from pET (Novagen), pQE (Qiagen), and pBAD (Invitrogen) vector backbones.

Tag Name	Tag Size	Tag Identity	Location	Solid Phase	Intact recovery	Recovery Condition	Supplier	Reusable Solid phase	K_d	Capacity (mg/mL)
His-tag	6aa	polyHistidine	N or C	Ni-NTA	Imidazole	Mild	Many	Yes		5-10
S-tag	15aa	RNaseS	N or C	S-protein Agarose	Urea, thiocyanates, $MgCl_2$, citrate	Partially denaturing	Novagen	Yes	~1 nM	0.5
MBP	~40 Kda	Maltose binding protein	N	Amylase resin	Maltose	Mild	NEB	Yes		2-4

DETECTION AND QUANTITATION WITH MULTIPLE TAGS

E. coli BL21-AI cells containing expression vectors encoding various multiple tag fusion proteins were grown to mid log phase and induced with IPTG and two different levels of arabinose.

A. Denaturing electrophoresis of soluble and pellet fractions with Coomassie Blue staining. B. Samples as shown in A treated with His-Tag stain. C, S-Tag detection analysis comparing the fluorescence obtained in the total cell lysate and in the soluble fraction.



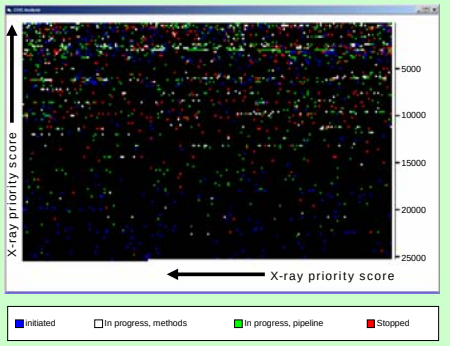
STATUS OF *ARABIDOPSIS* ORFS TARGETED BY CESG

Currently, ~1600 *Arabidopsis* ORFs have been targeted for investigation by CESG. Among these, ~1000 have been amplified by PCR, ~500 have been assembled into an expression vector, ~200 have been tested for expression, ~70 have given favorable solubility, ~30 have entered the purification stream, and 14 are currently in NMR or crystallization trials. The 367 "stopped" actions indicate a termination of workflow on an ORF due to a negative result, which may currently include decisions arising from the methods development process. The percentage of success reported for each action was calculated from the total number of ORFs tested in each category.

Status	# ORFs	Percentage	Status	# ORFs	Percentage
Selected targets	1627		Purified -	10	39
PCR +	979	75	Crystal +	4	57
PCR -	323	25	Crystal -	3	43
Cloned +	530	76	HSQC +	2	33
Cloned -	164	24	HSQC -	4	67
Expressed +	130	66	Quality X +	1	100
Expressed -	66	34	Quality X -	0	0
Soluble +	68	57	Structure +	0	
Soluble -	51	43	Structure -	0	
Purified +	16	61	Stopped	367	

DISTRIBUTION OF CESG TARGETS

This map represents the approximately 25,000 ORFs predicted for *Arabidopsis thaliana*, arranged by priority of analysis by x-ray crystallography (determined by a number of factors, see **http://uwstructuralgenomics.org/target/target.htm**). Those targeted by CESG thus far are shown with status indicated by color.



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. We are using both recombinational and restriction enzyme/ligation methods for gene cloning. Currently, ~1600 ORFs have been targeted by CESG. Greater than 95% of targeted intronless ORFs from genomic DNA of intronless genes and ~ 60% of intron-containing targets have been successfully amplified. Of these ~1000 ORFs, ~70 soluble protein products been advanced to more detailed purification and screening efforts. Future plans revolve around increasing the availability of protein targets by systematization of the protocols described here.

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