Center for Eukaryotic Structural Genomics

Protein Structure Initiative

CESG PSI-2 Progress Report Highlights PSI-2

Brian G. Fox, George N. Phillips, Jr., John L. Markley

University of Wisconsin-Madison, Department of Biochemistry, Madison, Wisconsin, USA http://www.uwstructuralgenomics.org

Abstract

CESG operates a full pipeline leading from target selection through structure determination and data deposition. As a technology development center in PSI-2, CESG's strategy is 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 success rates and further lower the costs of determining structures of eukaryotic proteins. CESG collects fine-grained information about individual pipeline steps within the single database created by its laboratory information management system (Sesame) and uses this information to identify critical steps for improvement and to evaluate new strategies for improving their efficiency. CESG has the ability to critically compare success rates of targets chosen to enlarge sequence-structure space (60% of our effort) with those chosen as being of biomedical relevance (20% effort) or nominated by the scientific community (20% effort). CESG's experience with eukaryotic targets provides ample background for investigations of proteins of biomedical relevance from humans and other vertebrates. CESG has developed bioinformatics methods for identifying targets of biomedical relevance and has used these in PSI-2 to launch two medically-relevant "workgroups" (groups of ~96 targets). In addition, CESG is collaborating with Dr. James Thomson (University of Wisconsin-Madison Medical School) to solve structures of proteins he and others identify as potentially playing key roles in the differentiation of human embryonic stem cells. To date, CESG has launched 54 stem cell targets through the cell-based pipeline and 48 through the cell-free pipeline. CESG has launched two sequence-structure space workgroups from eukaryotic thermophiles to complement three sequence-structure space workgroups from more familiar eukaryotes. Much of our research in the past year has focused on developing a robust platform that enables us 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. We are exploring approaches to the use of smaller quantities of protein for crystallization trials. In NMR, we have developed probabilistic tools for fast data collection and for automated processing and analysis of NMR data. Sesame, CESG's laboratory information management system, has undergone steady development. Over the past year, we have been working to make the data collected by Sesame fully compatible with the requirements of the PepcDB, have developed and deployed Sesame control of crystallization screening, and have created a web-based tool linked to Sesame to manage the growing number of structure requests CESG receives from the community. CESG is actively preparing for the transfer of vectors and clones to the PSI Materials Repository.

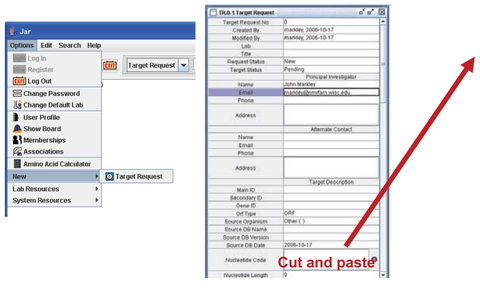
CESG Has a Well-Developed and Evolving LIMS: Sesame

- Available to Centers and individual investigators
- Single Database with Specialized Views: ORF, protein, solution, crystal, NMR, small molecule, structure deposition, sources (vendors), target submission
- Customizable: Laboratory resources, protocols, "actions"
- Captures Protocols and Results: Gel scans; NMR, MS, UV-VIS data; text
- Intuitive Query System
- Report Generation: XML, Excel input, Target DB, and PepcDB
- Supports Multiple Relational Database Management Systems: PostgreSQL 8+,
- Microsoft SQL Server 2005, or Oracle
- Tools for Managing Collaborative Projects
- Tools for Managing Shared Instruments

Sesame Jar Module: Web-Based Utility for External Structure Requests

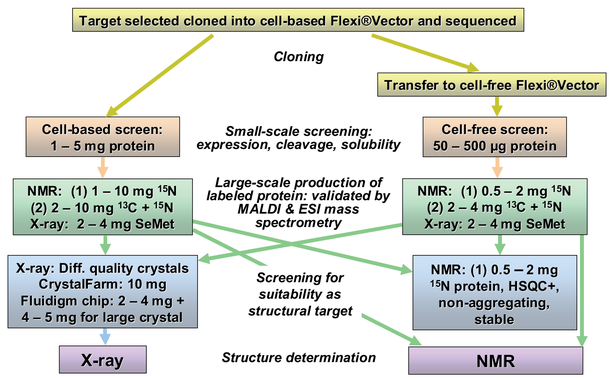
CESG's Online Request System enables scientists outside of CESG to suggest specific open reading frames (ORFs) to be worked upon by the Center, or submit recombinant protein for use in initial trials. Features of the online system include:

- Written in Java 2 and is Web-Based.
- Minimum computer requirements.
- Standardized request format.
- Save partial "Target Request" records and return to it later.
- Easily make multiple requests by using "Save As New".
- Sends automatic emails when milestones are achieved, e.g., "Cloned," "Expressed," "Purified," "Structure Solved".
- Allows naming of an alternate contact who will also receive auto emails.
- Submitters receive email when target is accepted, rejected, or work stopped.
- Submitters can check the status of any of their requests at any time.

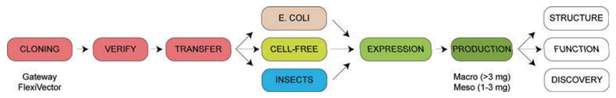


www.sesame.wisc.edu

Redundancy in CESG's Pipeline Leads to Higher Target Success

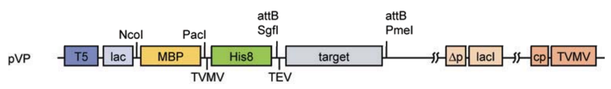


CESG Has Developed an Efficient Cloning Strategy for Protein Production in Multiple Platforms



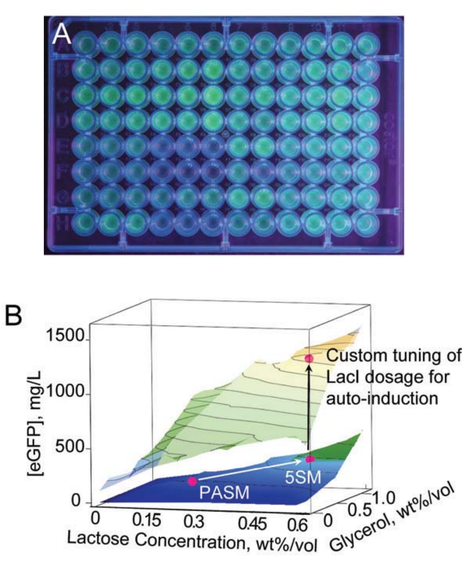
- Integrated cell-based and cell-free expression pipeline.
- Expression screening in both *E. coli* and wheat germ extract.
- An insect expression branch is in development.
- Decision tree for scale-up process is being developed by analysis of data captured into Sesame.

Modular Vector Design



- Essential elements are bounded by unique restriction sites.
- Promoter, solubility tags, affinity tags, linker combinations
- Over 70 variants; production vectors in NIH Materials Repository
- Target with N-terminal Ser cloning artifact was found to perform best.
- An Ala-Ile-Ala-target developed to minimize cloning costs failed in structure determinations and has been abandoned.
- Fusion proteins have been designed with *in vivo* proteolysis sites.
- Vectors have been designed to work in auto-induction media.

Evolution of CESG's Culture Media



- (A) Expression performance followed by GFT fusion.
- (B) Factorial design identified favored compositions different from PASM.
- Expression performance has been linked strongly to the level of lac repressor (LacI) expressed from the backbone vector.
- Yields of expressed protein > 1 mg/mL bacterial culture are now commonly achieved.

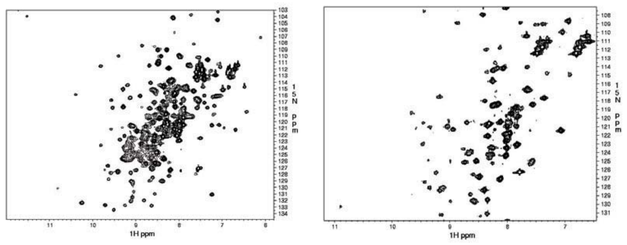
Simple Robotics for the Masses

Maxwell 16 System

Protemist DT II

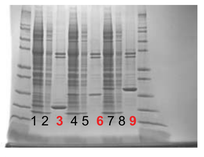
- (Left) Maxwell 16 system (Promega) supports automated parallel purification of 16 samples in 4 h.
- (Right) DT II benchtop robot (CellFree Sciences) supports automated transcription, translation, and IMAC purification of targets.
- These tools promise cost savings by providing small-scale methods for determining whether targets are likely to succeed in X-ray crystallization or HSQC NMR screens.

Samples Purified on the Maxwell 16



- (Left) ^{15}N HSQC NMR spectrum of His8-GFP. The protein sample was isolated from 8 mL of *E. coli* culture (self-cleaving VP 62K vector) grown at 35˚C and purified from 8 wells of the Maxwell. The yield was ~ 1.5 mg pure protein. The NMR sample at pH = 7.5 contained 0.2 mM GFP, 10 mM MOPS, 100 mM NaCl, 5mM DTT. NMR data collection took 1 h.
- (Right) ^{15}N HSQC NMR spectrum of a human embryonic stem cell target. The protein sample was isolated from 8 mL of *E. coli* culture (self-cleaving VP 62K vector) grown at 35°C and purified from 8 wells of Maxwell. The yield was ~ 0.25 mg pure protein. The NMR sample at pH 7.5 contained ~0.25 mM protein, 10 mM MOPS, 100 mM NaCl, 5 mM DTT. NMR data collection took 8 h.

Protein Purification on the DT II

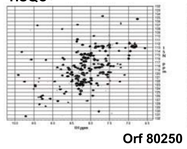


ORF80250: lane 1 – synthesis; lane 2 – flow through; lane 3 – purified soluble

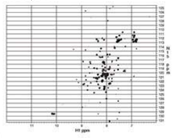
ORF80230: lane 4 – synthesis; lane 5 – flow through; lane 6 – purified soluble

ORF80229: lane 7 – synthesis; lane 8 – flow through; lane 9 – purified soluble

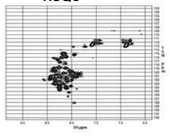
Orf 80229; 149 aa, w/His-tag 6 Pro; Peak No. 136/142; HSQC+



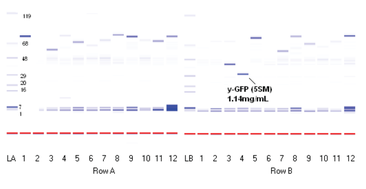
Orf 80230; 175 aa, w/His-tag 4 Pro; Peak No. 126/170; HSQC+/-



Orf 80250 HSQC-

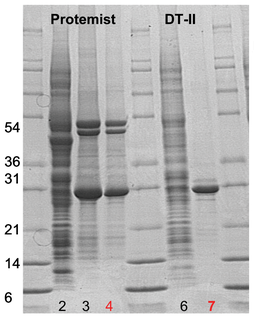


Goodbye Gels?



- (Left) Caliper LC-90 1 x Labchip can screen 96 targets in ~4 h ($225 / 96 targets). This compares 1-2 days for SDS-gels ($450 / 96 targets).
- (Right) The Caliper output can be made to look "gel-like."
- The Caliper unit provides accurate protein concentrations and molecular weights.
- The Caliper is most useful for assaying purified or partly-purified proteins.

Automated Cell-Free Preparation of Samples for Crystallographic Screening



ORF80717: 29 kDa (dimer by gel filtration)

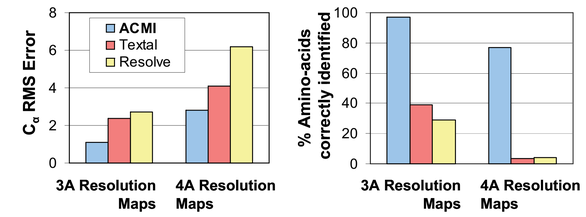
Lane
2: Protemist synthesis soluble; WG w/o AA
3: Protemist IMAC soluble
4: Protemist GF soluble
100 % Se-Met by ESI

6: DT-II synthesis soluble; WG-H w/o AA
7: DT-II robot IMAC soluble
100 % Se-Met by ESI
yield: 1.6 mg purified protein from 6 × 6 ml reactions

Nearly automated DNA to X-ray screening of Se-Met protein (only manual steps are manual buffer exchange and concentrations)

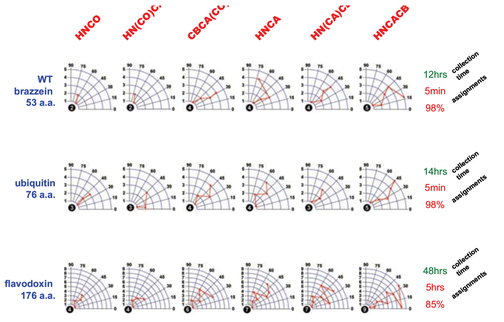
- "WG" is the standard wheat germ extract from CFS.
- "WG-H" is a new wheat germ extract from CFS that does not contain proteins that bind to the IMAC column (two large bands ~ 54 kDa).
- Level of Se-Met incorporation is close to 100%.

Automated Crystallography Map Interpretation (ACMI)



- ACMI approach shows promise for maps at low resolution.
- Published: DiMaio F, Shavlik J, Phillips GN (2006) *Bioinformatics* 22, e81.

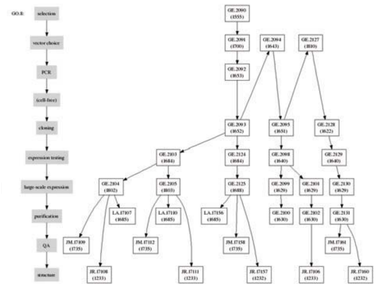
Toward Combined Fast NMR Data Collection and Automated Analysis



- Fast data collection by HIFI (adaptive, tilted-plane 2D collection of 3D NMR data) has been interfaced with PINE, which provides automated assignments and secondary structure determination.
- This represents a significant step toward our goal of automated, probabilistic NMR structure determination.

Reporting to PepcDB

- Reporting ***7751*** targets with ***68*** protocols.
- Last year was spent implementing atomized protocols, retraining and recoding PepcDB reporting.
- ~300 data errors remain.
- Project-wide training sessions.
- Developed a graphical tool for visualizing database linkages for each target.
- Leverages relationship between a type of directed graph and XML structure.



CESG is supported by the National Institute of General Medical Sciences through the Protein Structure Initiative grant number U54 GM074901.
Work also supported by grants P41 RR02301 (NMRFAM) and NLM 5K22LM8992 (HRE).