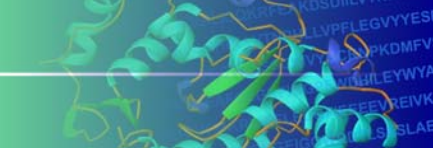
Center for Eukaryotic Structural Genomics

Protein Structure Initiative

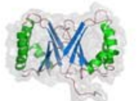


Pipeline strategy for efficient protein structure determination by NMR

Francis C. Peterson[1], Betsy L. Lytle[1], Jeanette Waltner[1], Sujatha Sampath[1], Robert Tyler[2], Jikui Song[2],
Claudia C. Cornilescu[2], Shanteri Singh[2], Qin Zhao[2], John L. Markley[2], and Brian F. Volkman[1]

[1]Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin, USA, 53226 and [2]University of Wisconsin-Madison, 433 Babcock Drive, Madison, Wisconsin, USA, 53706-1549, http://www.uwstructuralgenomics.org

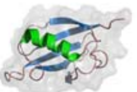
1JXC **At2g43510.1**
Pfam cluster: Toxin_3
Function: protease inhibitor



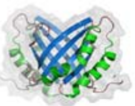
1Q53 **At3g17210.1**
Pfam cluster: B_3438
Function: unknown

1SE9 **At3g01050.1**
Pfam cluster: B_3438
Function: unknown

1T0G **At2g24940.1**
Pfam cluster: Cyt-b5
Function: steroid binding



1T0Y ***Ce*5O73**
Pfam cluster: B_9516
Function: tubulin cofactor B

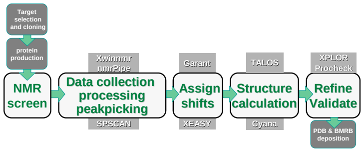


1RJJ **At5g22580.1**
Pfam cluster: B_3438
Function: unknown

1TIZ **At3g03410.1**
Pfam cluster: EF hand
Function: unknown

Abstract

The Center for Eukaryotic Structural Genomics (CESG) seeks to populate protein fold space through improvements in technology. To this end, we have developed a streamlined approach that employs cryogenic probes and automated analysis to more efficiently determine three-dimensional protein structures by NMR. A standardized experimental scheme eliminates the requirement for PhD-level instrument operators, and opens up the possibility for highly automated data collection. This highly modular pipeline strategy relies on a predefined framework for data management, which enables an efficient workflow even when multiple personnel participate at various stages of the process. The self-contained nature of individual steps allows for substitution of improved software tools or methods with advances in technology. The NMR structures determined in our laboratory demonstrate the effectiveness of our approach for proteins produced both in cell-based *E. coli* expression systems as well as the cell-free wheat germ system pioneered at the CESG.



Target selection and cloning

protein production

NMR screen

Xwinnmr nmrPipe

Data collection processing peakpicking

SPSCAN

Garant

Assign shifts

XEASY

TALOS

Structure calculation

Cyana

XPLOR Procheck

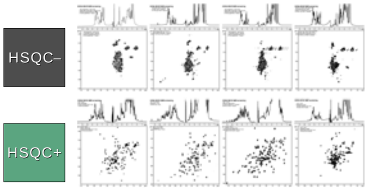
Refine Validate

PDB & BMRB deposition

CESG MCW NMR pipeline strategy

Sample screening by 2D NMR

Prospective targets for 3D structural analysis (<25 kDa, >0.5 mM protein) must first be tested for foldedness and sample stability under conditions used for 3D NMR data collection. Typically, NMR targets that pass cloning, soluble expression and purification stages of either the CESG *E. coli* cell-based or wheat germ cell-free production pipelines are produced in *U*-^{15}N labeled form and screened by 1D ^{1}H and 2D ^{1}H-^{15}N HSQC. NMR spectra are evaluated in terms of number, dispersion and uniformity of observed signals. HSQC+ targets are tested for stability over a ~1-2 week period, before *U*-^{15}N/^{13}C protein is requested from the production team.



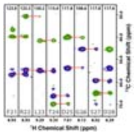
3D data acquisition on *U*-^{15}N/^{13}C protein

A standard set of 2D and 3D NMR experiments is acquired for each structure target, over a total data collection period of ~270h. All NMR data and acquisition parameters are captured by the Camel module of the Sesame database. Fourier transformation is performed using the NmrPipe program, and frequency-domain data are converted into the XEASY format. Signals in all 3D experiments are detected automatically and integrated using the SPSCAN program.

NOESY		Backbone		Sidechain	
Experiment	**Time**	**Experiment**	**Time**	**Experiment**	**Time**
^{15}N NOESY-HSQC	36 hr	^{15}N-^{1}H HSQC	0.3hr	3D CCONH	20hr
^{13}C ali NOESY-HSQC	36 hr	3D HNCO	7.5hr	3D HCCONH	14hr
^{13}C-^{1}H arom HSQC	1 hr	3D HNCOCA	7.5hr	3D HBHACONH	14hr
^{13}C arom NOESY-HSQC	36 hr	3D HNCA	7.5hr	HCCH-TOCSY	28hr
		3D HNCACB	14hr	^{13}C CT-HSQC ali	1.5hr
		3D HNCACO	14hr	^{13}C CT-HSQC arom	1.5hr
		^{15}N-^{1}H hetNOE	24hr		

Automated chemical shift assignment

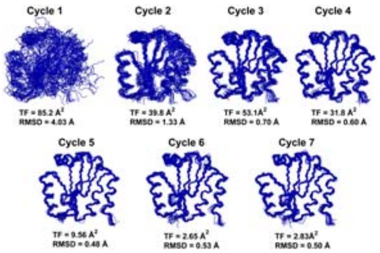
Peaklists obtained using the SPSCAN program are inspected in XEASY to eliminate spurious peaks or add unpicked signals. GARANT is used for automated assignment of backbone and sidechain chemical shifts. Peaklists containing tentative assignments are then inspected and edited manually to complete the assignment process. Results from GARANT calculations performed on 3D triple-resonance data from At3g51030.1 using different peakpicking strategies are compared in the table below.



Method/Atom	Expected	Correct	Wrong	Unassigned	% Correct
Manual	*Pick all 3D spectra manually in XEASY using 2D HSQC as reference*				
HN/N	84	79	3	2	94.0
CA	94	88	2	4	93.6
CB	75	64	8	3	85.3
CO	94	82	8	4	87.2
SPSCAN	*Pick all 3D spectra automatically using 3D HNCO as reference*				
HN/N	84	75	7	2	89.3
CA	94	82	8	4	87.2
CB	75	33	39	3	44.0
CO	94	82	8	4	87.2
Combined	*Automatic SPSCAN picking followed by manual editing in XEASY*				
HN/N	84	78	4	2	92.9
CA	94	81	9	4	86.2
CB	75	62	10	3	82.7
CO	94	82	8	4	87.2

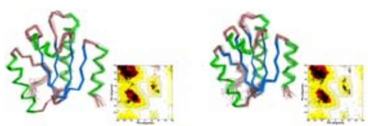
Automated structure calculations

Backbone torsion angle restraints are predicted from chemical shift values using the TALOS package and included in the initial round of structure calculations. Initial protein structures are generated using an iterative and fully automatic methodology for assignment of NOESY cross peaks provided by the CANDID module of CYANA.



Water refinement and validation

The final stages of structure refinement are accomplished through manual optimization of NOE assignments, NOE intensity-to-distance calibration functions, and backbone torsion angle restraints. Torsion angle dynamics structures that meet a set of objective criteria for agreement with experimental constraints and coordinate precision are subjected to molecular dynamics calculations in explicit solvent using the XPLOR-NIH package before final validation and deposition in the PDB and BMRB databases.



At3g51030.1

Cyana TAD ensemble

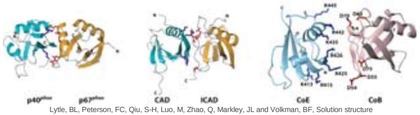
Ramachandran	
Most favored	86.3%
Also allowed	12.0%
Gen. Allowed	0.8%
Disallowed	0.8%

XPLOR water refined ensemble

Ramachandran	
Most favored	89.0%
Also allowed	10.4%
Gen. Allowed	0.3%
Disallowed	0.2%

C. elegans tubulin cofactor B

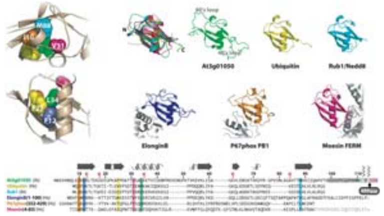
Proper folding and assembly of tubulin $\alpha\beta$-heterodimers involves a stepwise progression mediated by a group of protein cofactors A through E. Upon release of the tubulin monomers from the chaperonin CCT, they are acted upon by each cofactor in the folding pathway through a unique combination of protein interaction domains. The NMR structure of the N-terminal domain of *Caenorhabditis elegans* CoB showed that it adopts the ubiquitin fold. Based on modeling of the homologous cofactor E ubiquitin-like domain, we hypothesize that cofactors B and E may associate via their beta-grasp domains in a manner analogous to the PB1 and CAD superfamily of protein interaction domains.



Lytle, BL, Peterson, FC, Qiu, S-H, Luo, M, Zhao, Q, Markley, JL and Volkman, BF, Solution structure of a ubiquitin-like domain from tubulin-binding cofactor B, (2004) *J Biol Chem,* **150** 46787-93.

At3g01050, a putative target for prenylation

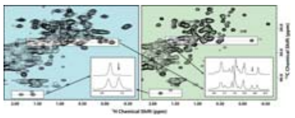
Because the β-grasp superfamily is so diverse, no obvious biological role can be inferred from the At3g01050.1 structure. However, the 10 proteins with strongest sequence similarity to At3g01050.1 are from plant species and are predicted to be prenylated. Prenylation has functional roles in plant cell cycle control, meristem development, and phytohormone signal transduction. Many prenylating enzymes recognize a C-terminal CAAX sequence, where C = Cys, A = aliphatic, and X is variable. At3g01050 possess a putative CAAX box motif and may be a target for post-translational modification by prenyltransferases. Interestingly, the At3g01050.1 C-terminus, CVCSVM, could support double lipid modification of the type required for the proper localization and function of Rab GTPases which are farnesylated at the cysteine of a CAAX box and subsequently palmitoylated at a second nearby cysteine.



Vinarov, DA, Lytle, BL, Peterson, FC, Tyler, E, Volkman, BF and Markley, JL Cell-free protein production and labeling protocol for NMR-based structural proteomics, (2004) *Nat Meth,* **1** 149-154.

Stereoarray Isotope Labeling (SAIL)

The ^{1}H density in a protein can be reduced by nearly 50% by incorporation of SAIL amino acids, synthesized in the laboratory of Prof. Kainosho (Metro. Univ. of Tokyo). ^{1}H-^{13}C HSQC spectra of *U*-^{15}N/^{13}C (left) and SAIL (right) At1g44790.1 produced in the wheat germ cell-free expression system illustrate the simplifying effects of this labeling strategy. SAIL labeling improves NMR sensitivity, reduces spectral degeneracy, and improves 3D structure determination for larger proteins by enabling stereospecific assignment of every NOE constraint to a single proton.

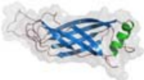


1TQ1 **At5g66040.1**
Pfam cluster: Rhodanese
Function: unknown

1XFL **At3g51030.1**
Pfam cluster: Thioredoxin
Function: redox protein

1XO3 **Mm202773**
Pfam cluster: B_45334,B_14998
Function: unknown

1XO4 **At1g77540.1**
Pfam cluster: B_34288,B_2135
Function: acetyl transferase



1XO8 **At1g01470.1**
Pfam cluster: LEA_2
Function: unknown

1XO9 **At3g03773.1**
Pfam cluster: B_4851
Function: unknown

1XOY **At3g04780.1**
Pfam cluster: DUF1000
Function: unknown

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