Protein production: cell-based vs. cell-free

John Markley
University of Wisconsin-Madison
markley@nmrfam.wisc.edu

Information on protocols at:
http://uwstructuralgenomics.org

Presented at the AMMRL meeting at 2008 ENC

Protein production currently is the major bottleneck in protein structure determination by NMR and X-ray

1) **Expression levels can be low: particularly for eukaryotic (vertebrate) proteins and membrane proteins**
2) **Protein production methods that work for unlabeled proteins (rich medium) may fail when labeling is attempted (minimal medium)**
3) **Proteins may be insoluble or misfolded**
4) **Soluble proteins may be natively disordered, partially folded, or aggregated**
5) **Well-folded and soluble proteins may be unstable**

Structural Biology vs Structural Genomics

Structural biology

- Frequently takes advantage of prior information about a system—including how to make the protein
- Many approaches are tried "once" – results are anecdotal
- Limited resources prevent trying "new" approaches (or even the bewildering array of those in the literature)

Structural genomics

- Procedures are tried repeatedly and generate valid statistics about what works (or doesn't)
- Provides the resources to try out new approaches and critically review those reported by others
- "Best practices" are emerging
- PSI is committed to sharing information and products

Resources supported by the NIH Protein Structure Initiative you should know about

PSI Knowledgebase: information hub linking PSI sites and other sites; software for protein structure prediction; access to protocols and technology

TargetDB: database of ORFs selected by SG groups around the world

PEPCdb: database of ORFs and the protocols that were used on them (successfully and not)

Materials Repository: plasmids and cloned genes generated by the PSI centers

Protein Production and Crystallography Workshop (PPCW): held yearly in Bethesda open to public—also covers NMR technology

Why do we use cell-free protein production?

- **Small scale screening is inexpensive ($3.50 / well in 96-well format), fast (DNA to gel scans in 2 days), and highly predictive of scale-up results**
- **Labeled proteins can be produced in two days or less, including purification (rapid because volumes are low)**
- **Higher supplies costs (wheat germ extract and labeled amino acids) are offset by lower labor costs**
- **Success rate with eukaryotic proteins including membrane proteins is higher than from *E. coli* cells**

R. Tyler et al. (2005) *Proteins* 59:633

PSI

Save time & money: clone sequencing; micro-, meso-, and macro-level protein production

- Sequence destination clones
- Screen first on a very small (micro) scale to find conditions that produce sufficient soluble protein
- Use an analytical (calibrated) gel filtration column to check for monodispersity and oligomerization
- Use mass spectrometry to check for correct protein and labeling
- Make ^{15}N-labeled protein on a meso scale for testing for folding and stability
- Only make double-labeled (^{13}C, ^{15}N) or triple-labeled (^{2}H, ^{13}C, ^{15}N) protein on a macro scale when conditions have been perfected with inexpensive [U-^{15}N]-protein

General approach: protein production for structural and functional genomics / structural biology

Micro-scale (2 – 5 μg). Screening of targets for expression and solubility: different approaches for water soluble and membrane proteins

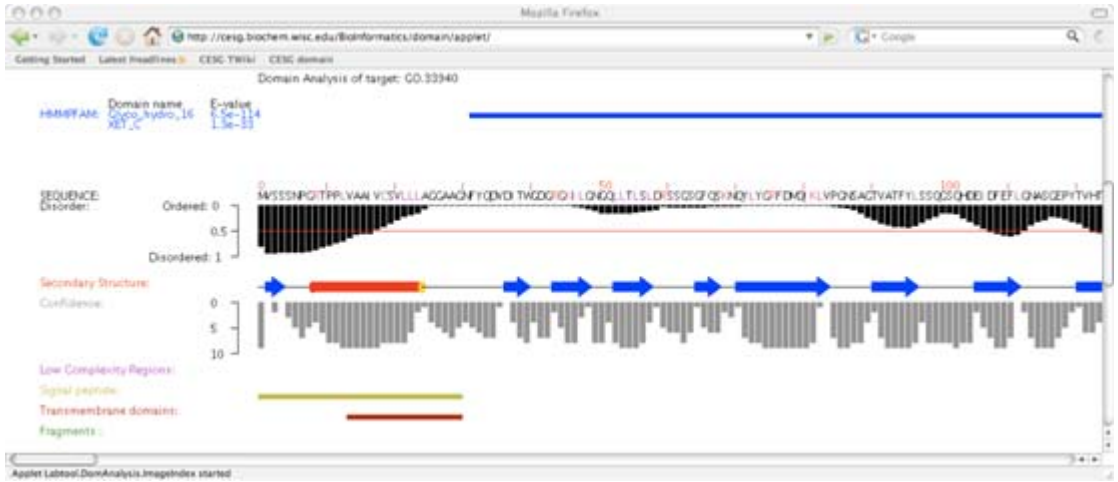
Meso-scale (0.5 – 1 mg). Production of labeled protein for biophysical and activity screening

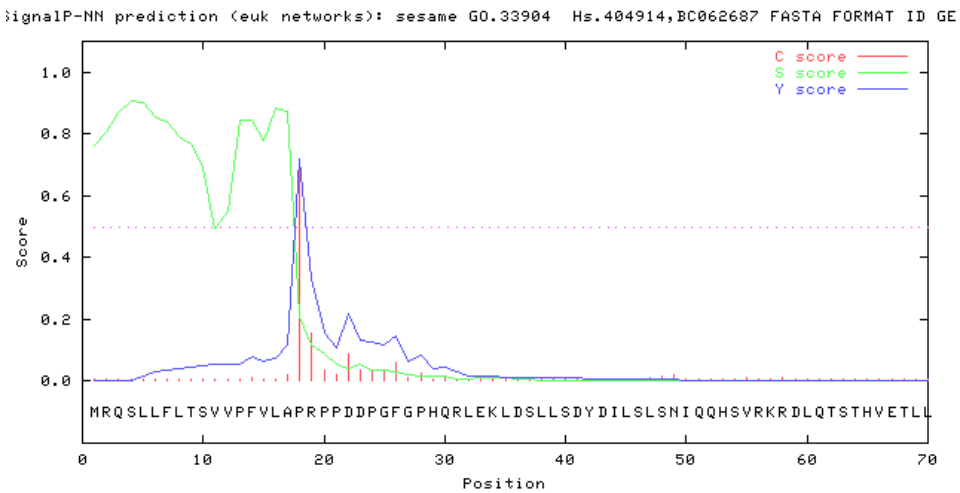
- Check for expected mass (ESI and MALDI mass spectrometry
- Analytical scale gel exclusion chromatography (monodispersity and oligomeric state)
- Analysis for crystallizability (Se-Met label)
- Analysis for suitability as NMR structural target (^{15}N label)
- Ligand binding assays (calorimetric, NMR, MS)
- Enzymatic assays

Macro-scale (2 – 10 mg). Production of labeled protein for structure determination by NMR (^{13}C+^{15}N label) or X-ray (Se-Met label)

Be forewarned: use bioinformatic tools for target selection and optimization

- OMIM (Online Mendelian Inheritance in Man) database used in identifying biomedical relevance
- Domain analysis
- Disorder analysis (PONDR)
- Predicted secondary structure
- Low complexity analysis
- Transmembrane segment prediction
- Signal sequence prediction





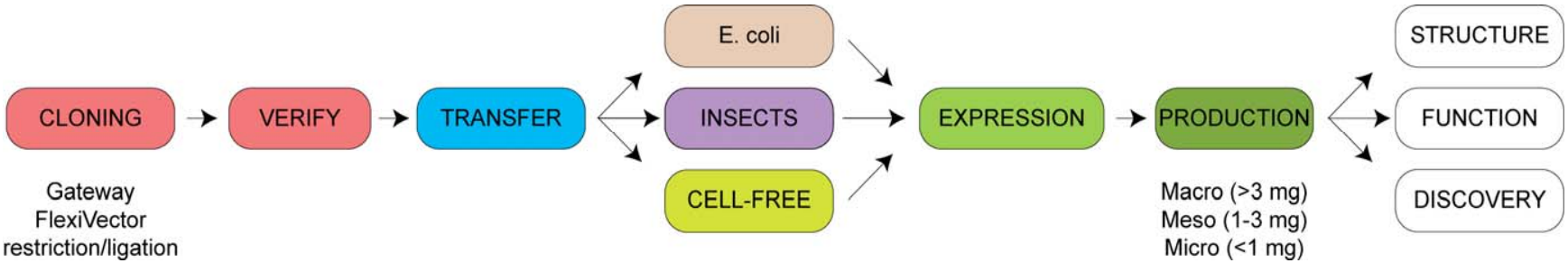
Discussion points: Tricks of the trade—what to do when

1) No expression
2) Expression but no solubility
3) Yield is low
4) Solubility but no cleavage of fusion protein
5) Results are great with rich medium but poor with minimal medium
6) Protein with multiple disulfide bonds fails to fold properly
7) No *E. coli* approach works
8) Neither *E. coli* cells nor cell-free works
9) Protein gives a poor 1D ^{1}H NMR or 2D ^{15}N HSQC
10) Protein looks good for a short while but crashes out later
11) Protein is too big or too dynamic for analysis by conventional ^{13}C, ^{15}N double labeling

As a technology development center CESG (our structural genomics project) is developing approaches:

1) To lower costs of cloning
2) To develop vectors that support cloning once with economical transfers into vectors that support a variety of expression alternatives
3) To develop low-cost small-scale screening methods that are predictive of scale-up
4) To develop auto-induction media for producing labeled proteins
5) To produce proteins by in vitro (cell-free) methods (both water soluble and membrane proteins)
6) To screen NMR samples for optimal solution conditions and to check construct for suitability: protein must be folded, monodisperse, and stable

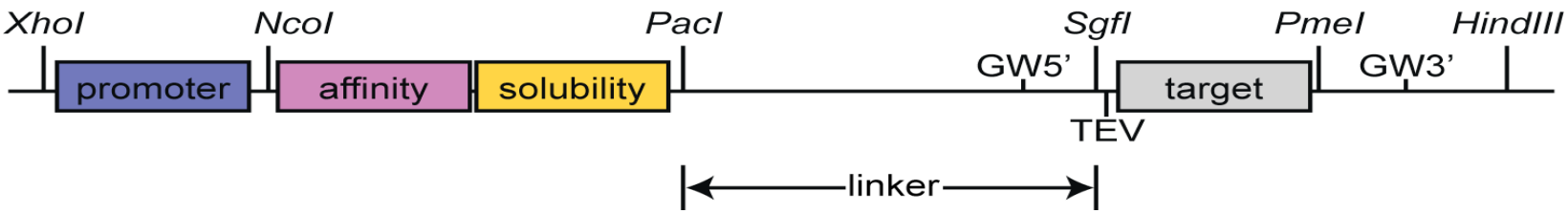
Unified cloning platform for cell-free and cell-based protein production



- **Cell-based and cell-free approaches complement one another and stimulate innovation**
- **FlexiVector has been chosen because the Gateway recombination site is inhibitory to cell-free translation whereas the FlexiVector cloning site is not**

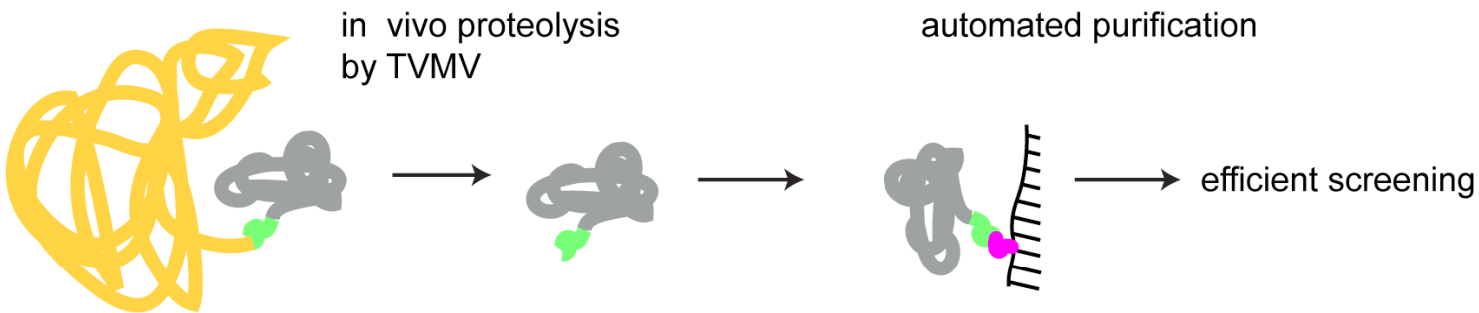
Blommel P, Fox B (2006) *Protein Expr Purif* 47:562

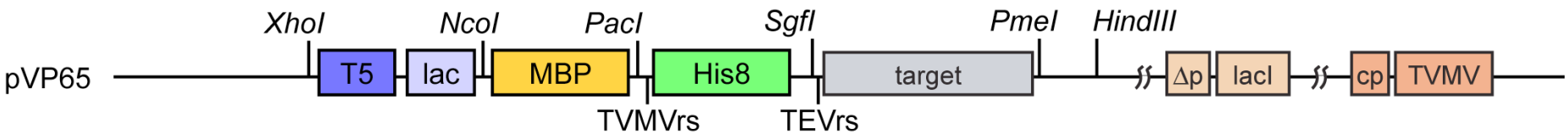
Flexi Vector design for cell-based protein production



- **CESG has produced >75 vector variants**
  - **All essential functionalities bounded by unique restriction sites**
  - **Reserved set is *XhoI, NcoI, PacI, SgfI, PmeI, HindIII, AvrII, BsiWI***
  - **Many combinations of promoter, affinity, solubility, and linker tags**
- **Production vectors will be sent to the PSI Materials Repository**
- **In vitro TEV proteolysis gives either S-target or AIA-target**
- **Novagen and Qiagen insect-cell free approaches were tested--but results were disappointing**
- **pFastBac being converted to Flexi Vector in SRA with Boehringer**

Vector design for cell-based protein production: addition of in vivo proteolysis





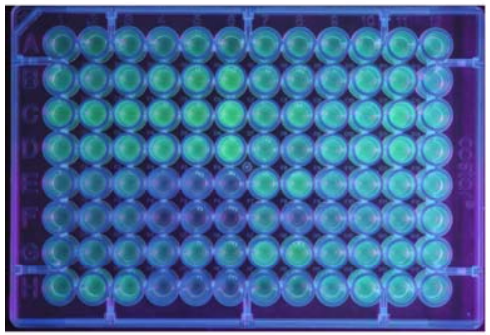
CESG carries out its **macro-scale** *E. coli* cell growths in polyethylene terephthalate vessels (pop bottles purchased from a soda bottling plant) and uses a self-induction medium

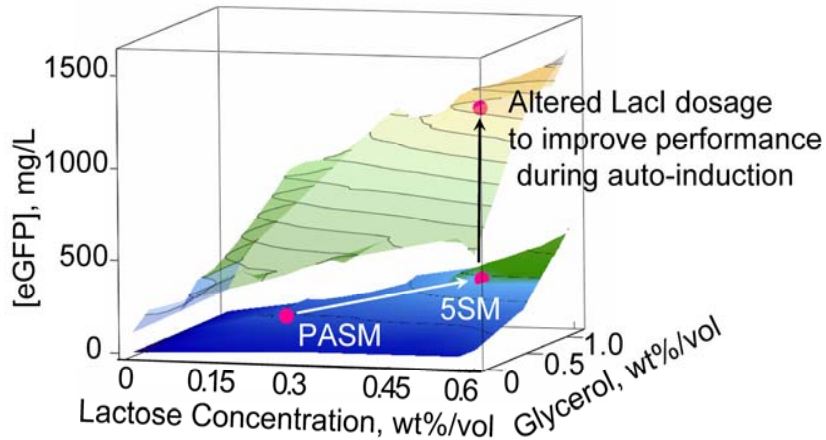
Yields and labeling efficiency (Se-Met and ^{15}N/ ^{13}C) have beem improved through optimization of self-induction media

H. K. Sreenath et al. (2005) *Protein Expr Purif* 40:256

R. L. Tyler et al. (2005) *Protein Expr Purif* 40:268

An evolved auto-induction medium





- **PASM is the original self-induction medium described by studier**
- **5SM is the medium we introduced for cell free on the basis of factorial design**
- **Further analysis showed that yields could be improved by lowering the dosage of the Lac inhibitor (LacI)**

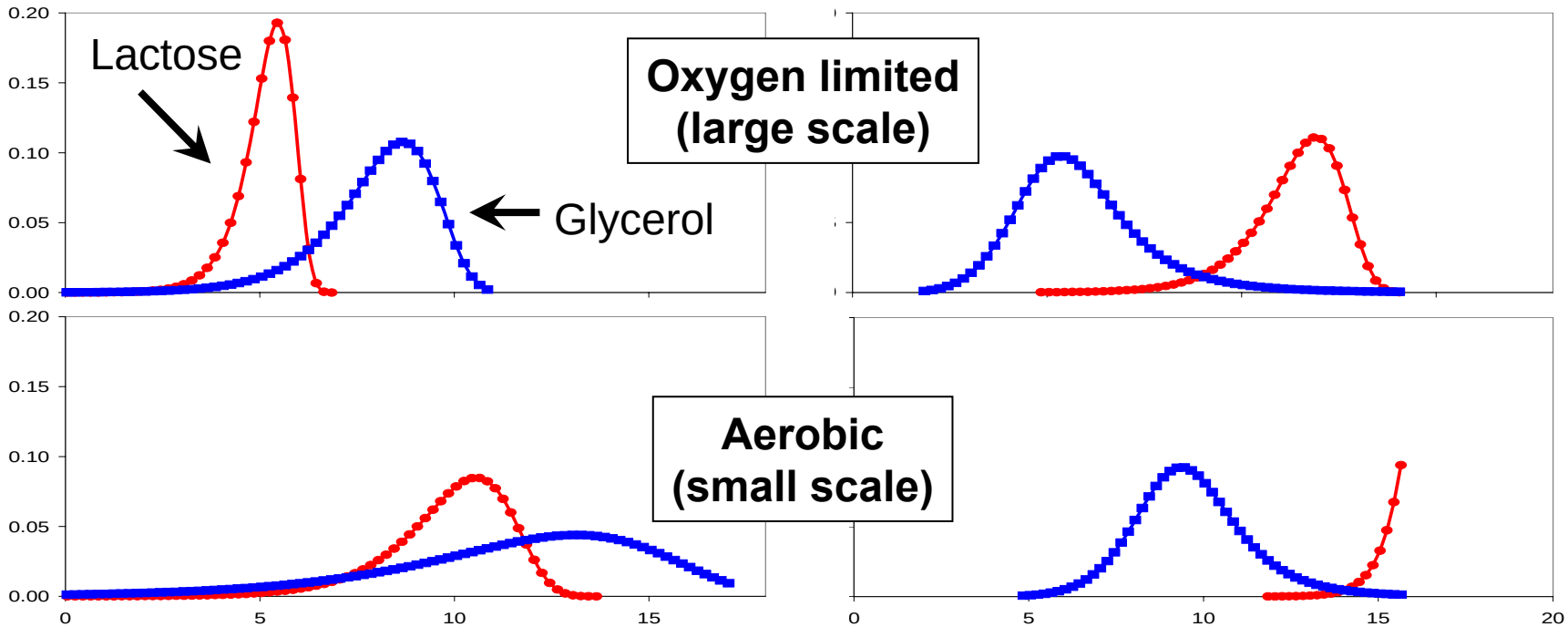
Blommel, Becker, Duvnjak and Fox (2007) Biotech Prog, 23:585

Experiments have shown why auto-induction media fail and how the problem can be remedied

Glucose is utilized first, but the fate of lactose and glycerol depends on conditions

T7 – No additional *LacI*

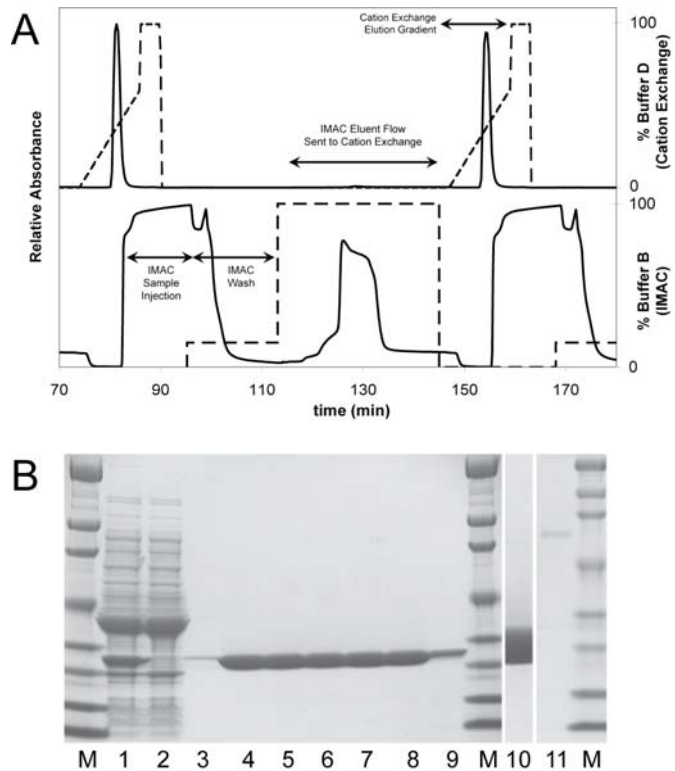
T5/*Lac*^{2} + 200x wt *LacI*



Blommel, Becker, Duvnjak and Fox (2007) Biotech Prog, 23:585

Application of the improved self-induction medium to the production of engineered His-tagged TEV protease

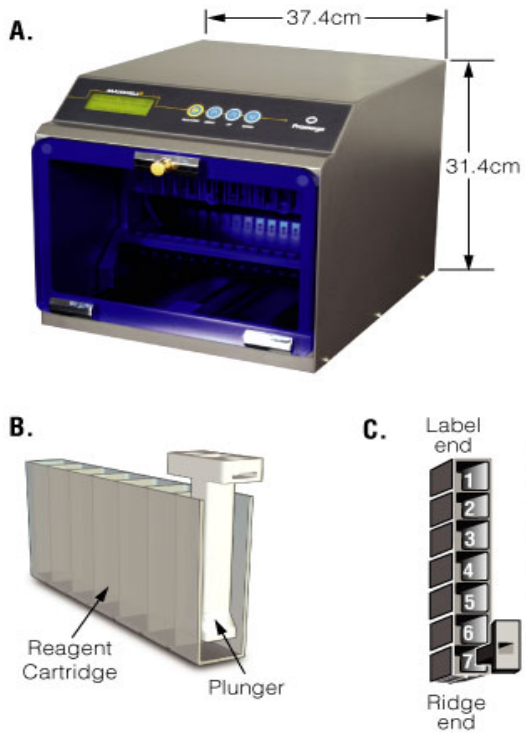
- TEV protease is an expensive reagent used by CESG
- 400 mg pure protease produced from 1 L of culture medium (15 mg / g cells)
- One of CESG's most popular 'products': 13 MTA requests to date
- TEV-Δ238 will be deposited in the PSI Materials Repository



Blommel and Fox (2007) Protein Expr Purif, 55:55

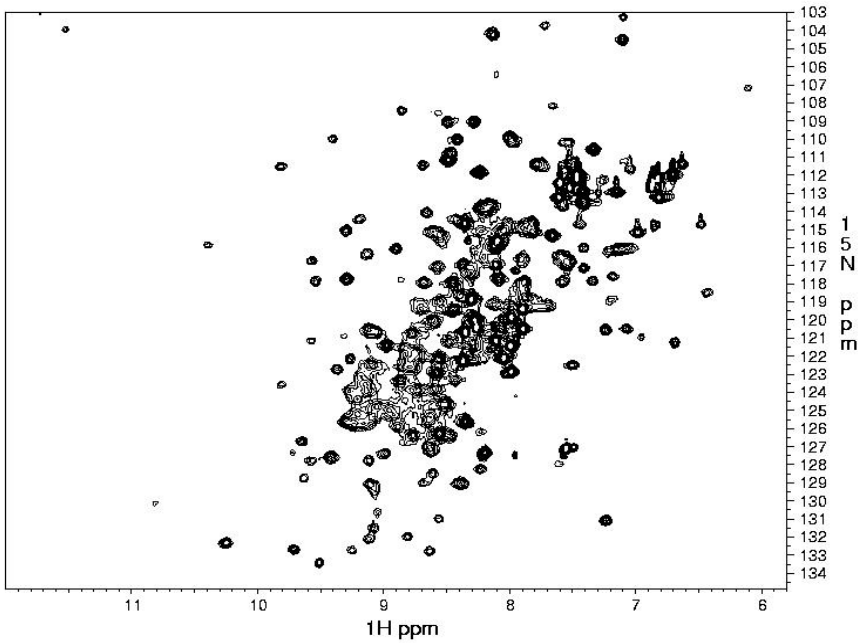
Maxwell-16: inexpensive bench-top robotic system

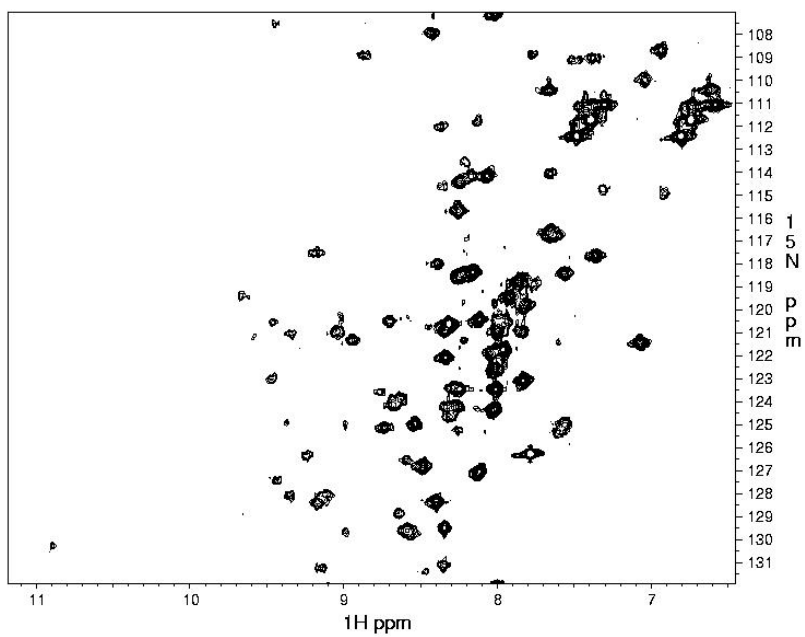
- Collaboration with Promega
- Cells to IMAC purified protein in 40 min
- 16 samples processed at one time
- Rapid information on protein production and solubility
- Can combine wells to produce sufficient [^{15}N]-protein for NMR or SeMet-protein for crystallization trials.



R. Frederick, et al. J Struct Funct Genomics, web (2008)

Maxwell-16 production of [^{15}N]-samples for NMR screening from 8-mL *E. coli* cultures



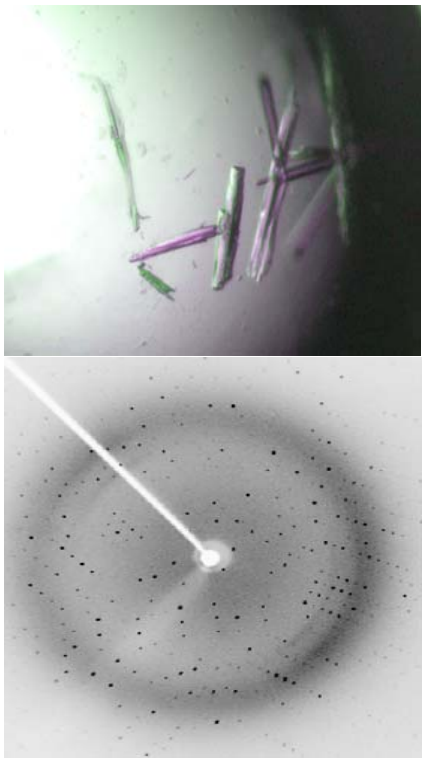


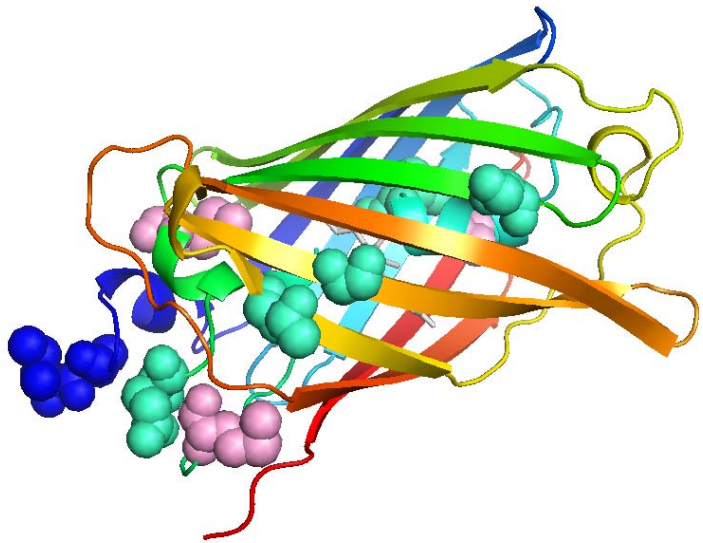
- **^{15}N HSQC NMR of His$_8$-GFP expressed in self-cleaving vector pVP62K at 35 ˚C. The NMR acquisition time: 1 h.**
- **1.5 mg purified from 8 wells of Maxwell in 40 minutes.**
- **Cost is $50 (labeled medium and purification cartridges). 40 minute run (1.5 mg from 8 mL).**

- **^{15}N HSQC NMR of human embryonic stem cell protein expressed in self-cleaving vector pVP62K. The NMR acquisitions time: 8 h.**
- **0.25 mg protein purified from 8 wells of a Maxwell-16**

***R. Frederick, et al. J Struct Funct Genomics, web* (2008)**

Maxwell-16 purification of an engineered synthetic GFP and structure determination by X-ray crystallography





R. Frederick, et al. J Struct Funct Genomics, web (2008)

Residues Ile and Ala from the N-terminal are shown in blue spheres. The mutated residues included in the protein are shown as cyan spheres. Selenomethionine residues are shown in violet spheres. The chromophore is shown as grey sticks. Resolution of protein was 1.7 Å.

Comparison of protein production on CESG's wheat germ cell-free and E. coli cell-based platforms

96 targets (7 – 37 kDa) put through CESG's cell-based and cell-free platforms

Proteins were ^{15}N labeled if possible

Outcome was followed all the way through HSQC screening and NMR structure determination

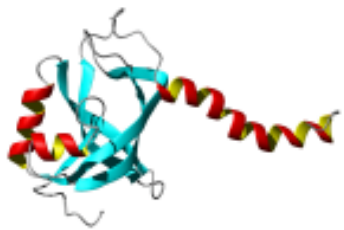
Wheat germ cell-free		*E. coli* cells	
HSQC +	HSQC +/–	HSQC +	HSQC +/–
10	1	5	1

R. Tyler et al. (2005) *Proteins* 59:633

PSI

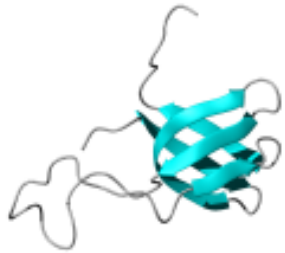
Example of NMR structures of proteins produced by CESG's cell-free pipeline

(1) Hs.78877
11 kDa
PDB: 2G2B

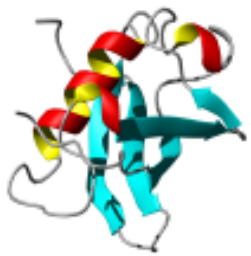


(2) At5g39720.1
19 kDa
PDB: 2G0Q

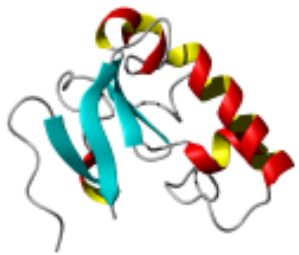
(3) At5g66040.1
14 kDa
PDB: 1TQ1



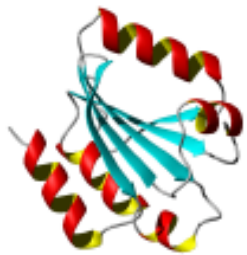
(4) Dr.13312
12 kDa
PDB: 2FB7



(5) At3g01050.1
13 kDa
PDB: 1SE9



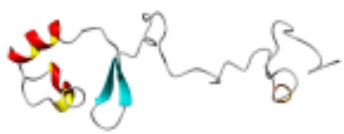
(6) At2g24940.1
11 kDa
PDB: 1T0G



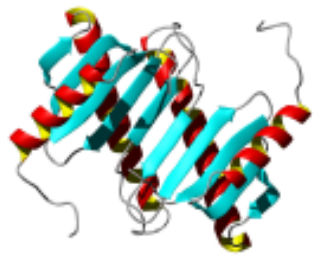
(7) At3g51030.1
14 kDa
PDB: 1XFL

(8) Hs.102419
13 kDa
PDB: 1ZR9

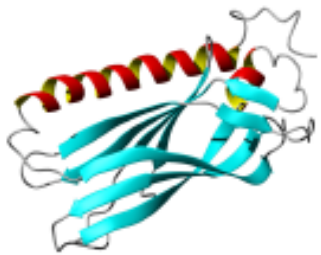
(9) Hs.157607
14 kDa
PDB: 2ETT



(10) At2g23090.1
9 kDa
PDB: 1WVK



(11) P62627
dimer 22 kDa
PDB: 1Y4O



(12) At2g46140.1
19 kDa
PDB: 1YYC

How to get started with cell-free protein production

- **The learning curve can be steep**
- **Visit a lab that uses the method**
- **Attend a workshop**
- **Purchase a kit**

  Roche sells kits for *E. coli* cell-free

  Cell-free Sciences sells kits for wheat germ cell-free without amino acids

  Cost for 3-4 mg protein

  Cell-free extract: $160

  Labeled amino acids: $200

- **You can make your own *E. coli* cell free extract**

  T. Torizawa ….. M. Kainosho (2004)

  ***J. Biomol. NMR.* 30:311**

CellFree Sciences DT-II robotic system: DNA to pure protein in 35 h

Mode of operation:

- Six 6-mL samples
- Translation in bilayer mode: wheat germ extract and mRNA on bottom, amino acids, energy source on top
- Fully automated transcription, translation, and batch method affinity purification

Yield of purified protein:

- 0.1 – 0.3 mg per sample
- 0.6 – 1.8 mg from all six samples

Running time:

- 35 hours including purification

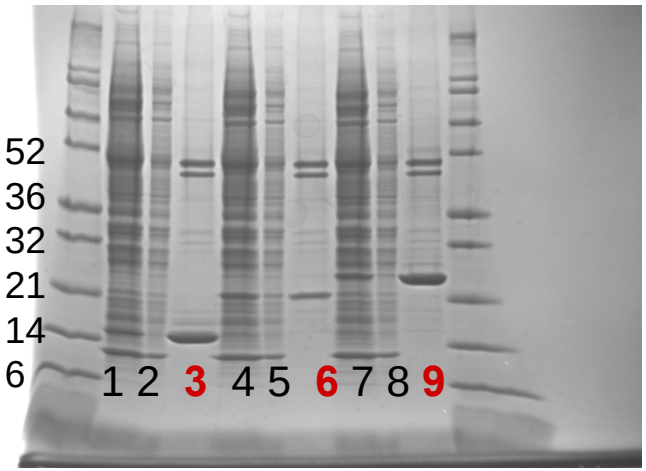
Cost:

- ~$80 per sample
- + $10 per sample for [^{15}N]-amino acids

Additional work required to prepare an NMR sample:

- Buffer exchange and concentration

Protein production and purification on DT-II for subsequent NMR analysis

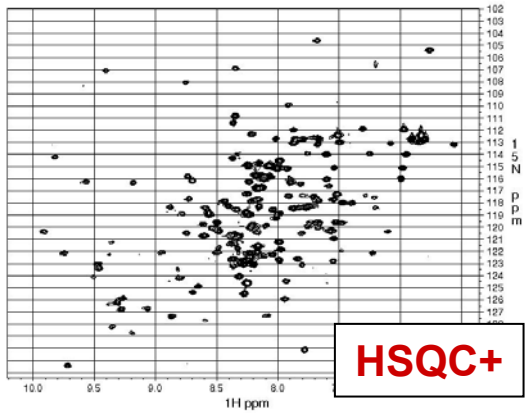


ORF80250: lane 1 – synthesis
lane 2 – flow through
lane 3 – purified soluble

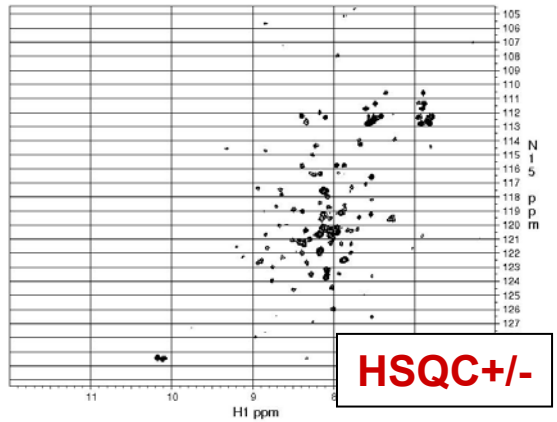
ORF80230: lane 4 – synthesis
lane 5 – flow through
lane 6 – purified soluble

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

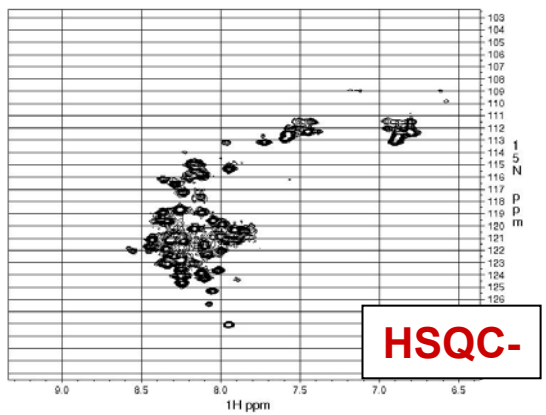
Orf 80229; 149 aa, 6 Pro, w/His-tag
Peak count 136/142



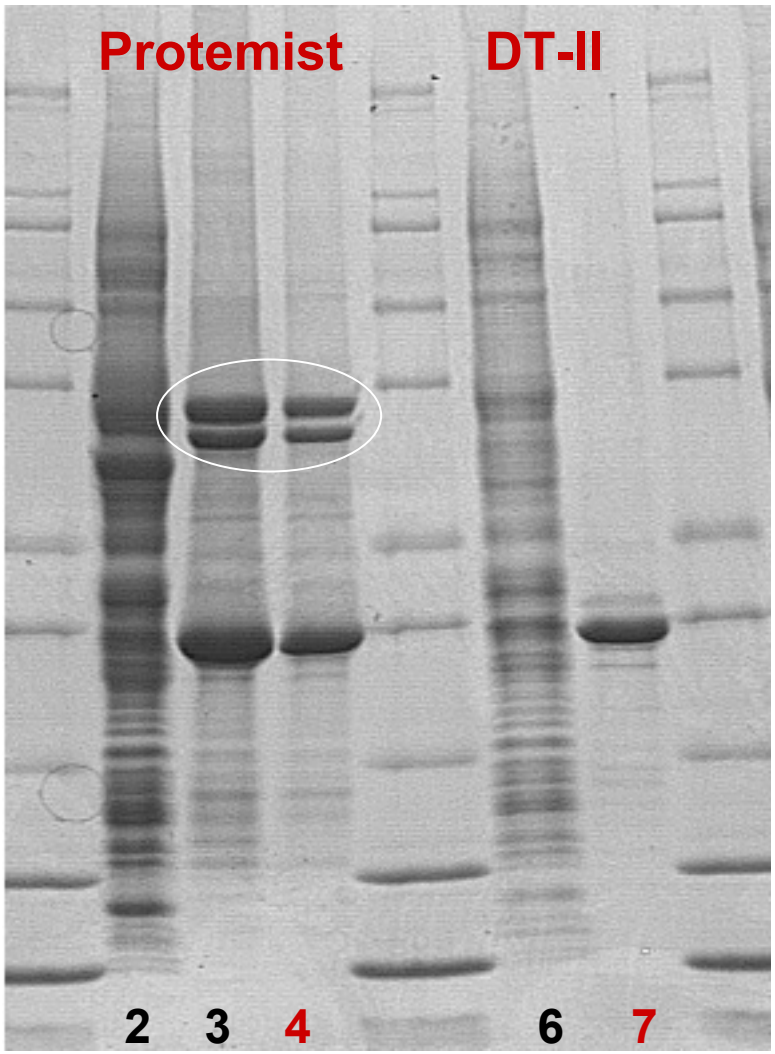
Orf 80230; 175 aa, 4 Pro, w/His-tag
Peak count 126/170;



Orf 80250



Use of a new wheat germ extract (WG-H) that is free from IMAC-binding contaminants (~54 kDa) found in the original extract (WG)

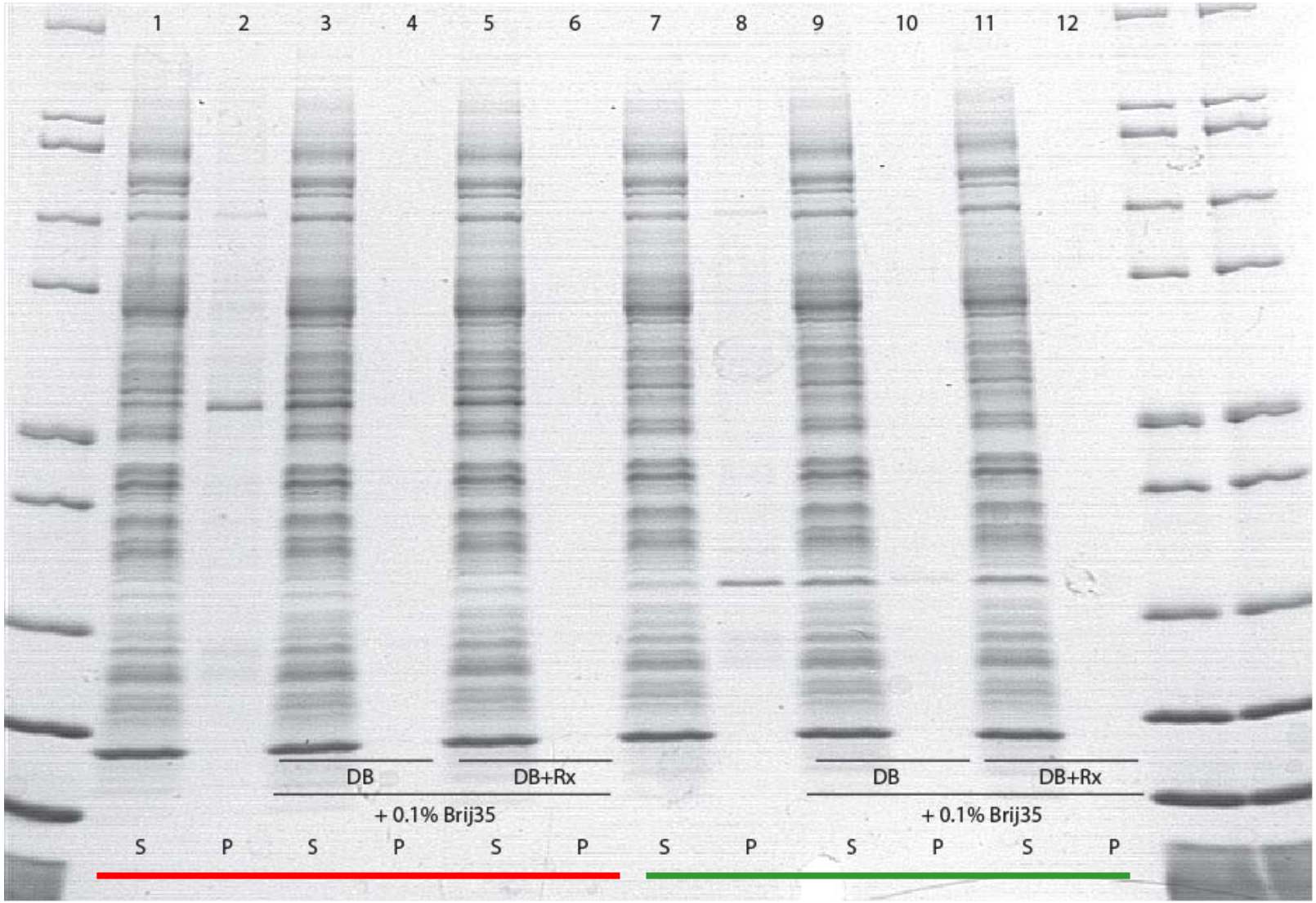


Production of a Se-Met target for X-ray: 58 kDa dimer by gel filtration; 29 kDa on SDS PAGE

Produced from extract WG
2: Soluble fraction
3: Soluble fraction after IMAC
4: Product after gel filtration step
100 % Se-Met by ESI

Produced from new extract WG-H
6: Soluble fraction
7: Soluble fraction after IMAC
100 % Se-Met by ESI
Yield: 1.6 mg purified protein from 6 x 6 ml reactions

Membrane protein targets synthesized on DT-II in 1.2-mL bi-layer mode



Lanes 1-6: human sterol Co-A desaturase

Lanes 7-12: ORF3093, *Arabidopsis* inner mitochondrial membrane protein

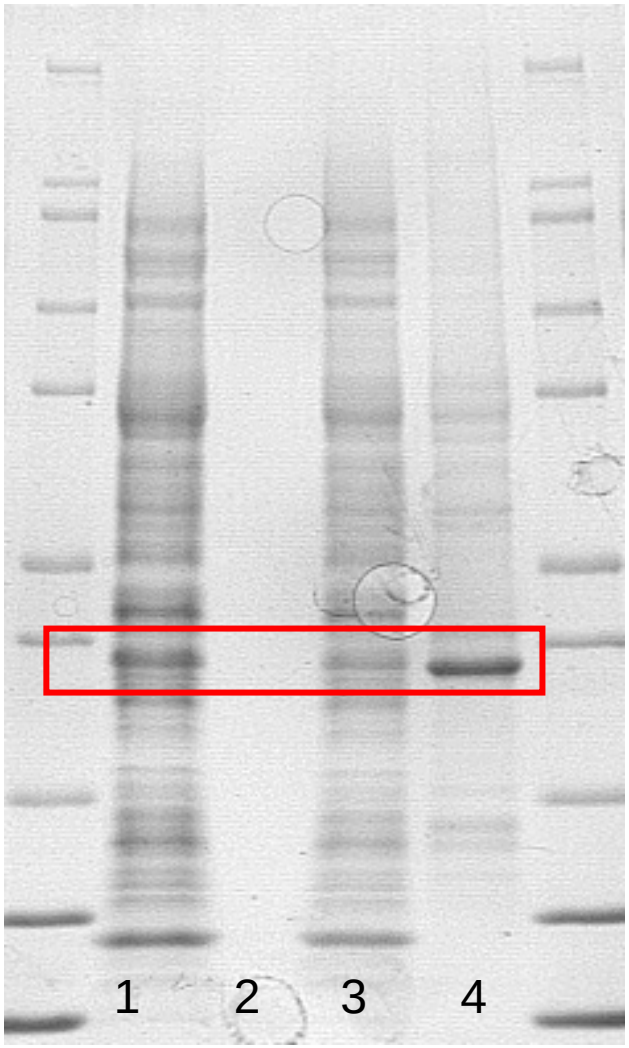
DB – detergent only in dialysis buffer; DB+Rx – detergent in dialysis buffer and reaction mix

Cell-free production of membrane proteins without detergent and in the automated DT-II bilayer system with detergent (Brij35)

ORF	Annotation	MW	PI	E w/o det	S w/o det	E w/ det	S w/ det
1255	membrane protein, putative	35,413	9.4	H	L	0	
4031	transmembrane like protein	24,156	6.3	H	L	H	H
5555	transmembrane protein, putative	24,346	6.2	M	L	H	H
14480	putative membrane protein	22,028	9.8	M	L	0	
2638	transmembrane like protein	24,665	6.1	H	L	M	H
10464	putative membrane transporter	60,171	9.6	H	L	H	H
10496	plasma membrane intrinsic protein 1a	30,689	9.1	M	L	H	H
16539	probable plasma membrane intrinsic protein 1c	31,603	9	M	L	H	H

H = high
M = medium
L = low

Production of a membrane protein in detergent and exchange into a detergent compatible with NMR



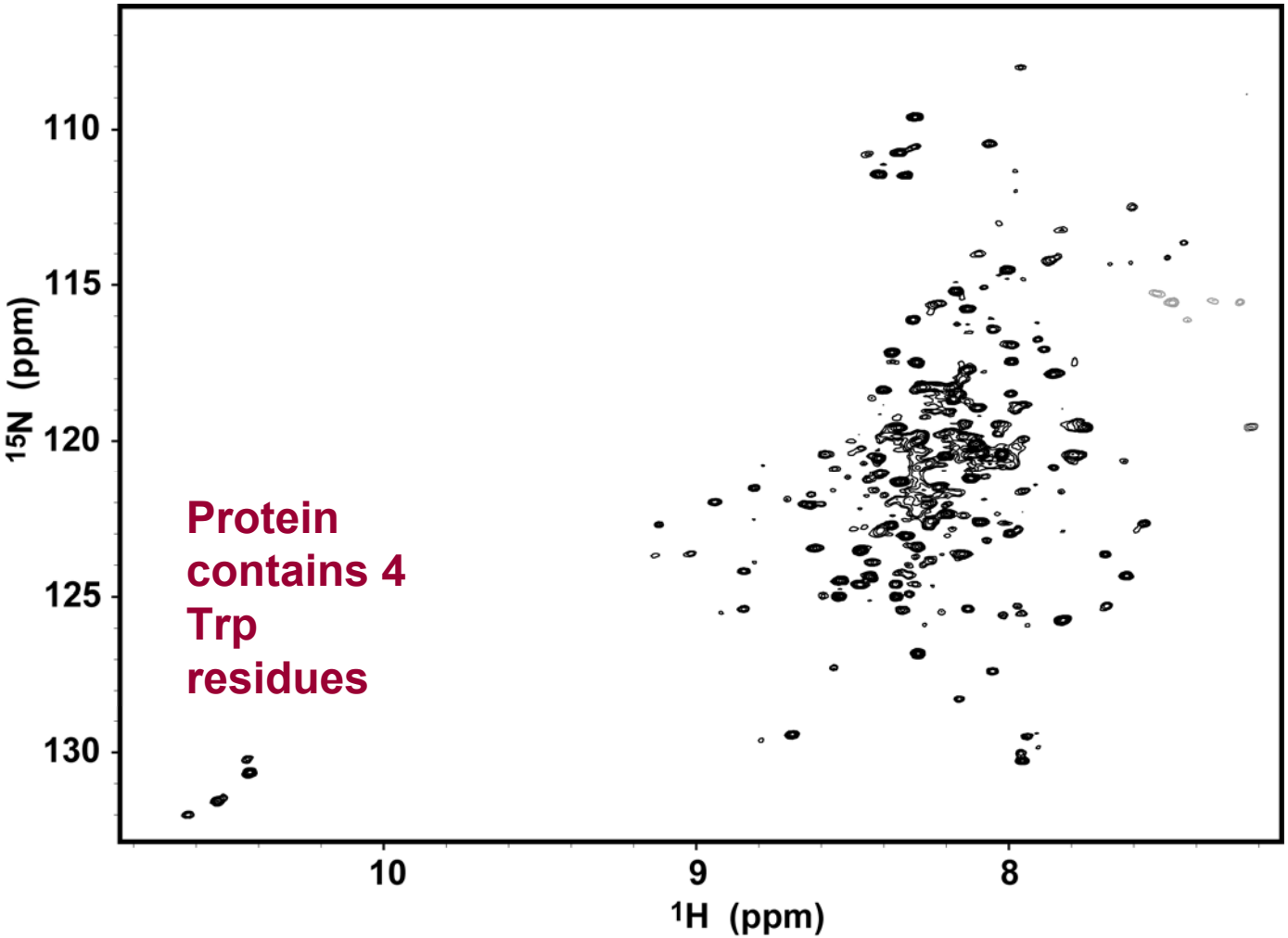
Arabidopsis thaliana **Columbia "transmembrane like" protein: 220 aa residues, pI 6.4**

Protein synthesis in the presence of 0.1% Brij35 (non-ionic detergent compatible with WG cell-free protein expression)

Purified protein was exchanged into 0.5% Fos-choline-12 (lipid-like zwitterionic detergent)

Lane 1: synthesis soluble
Lane 2: synthesis pellet
Lane 3: flow through IMAC
Lane 4: soluble after IMAC

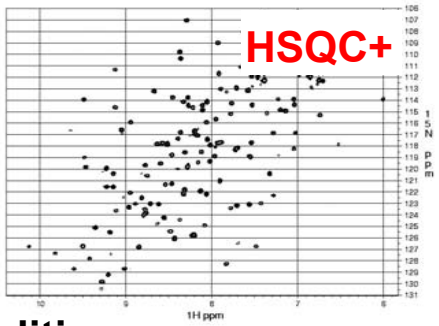
^{1}H-^{15}N TROSY spectrum of a membrane protein prepared by wheat germ cell-free in the presence of detergent and transferred into Fos-choline-12 micelles



At1g14010.1 (sample JR 18470) at 600 MHz field
"transmembrane-like protein" 24 kDa

Buffer optimization

Target:
At3g04780.1

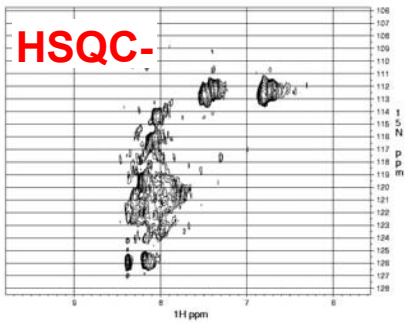


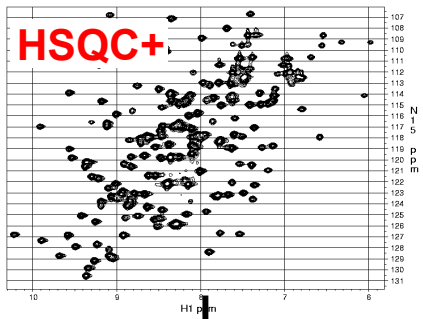
Initial conditions led to unfolding:
5 mM MES
50 mM NaCl
pH 6.0

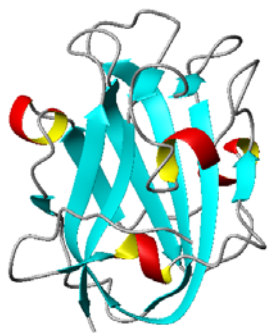
2 days

> 4 weeks

Best conditions:
10 mM DTT
50 mM KH2PO4
pH 7.2

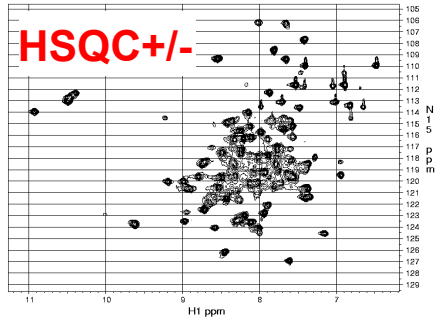


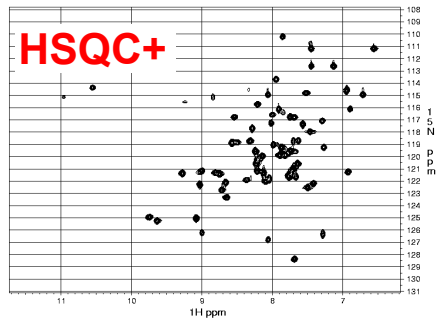


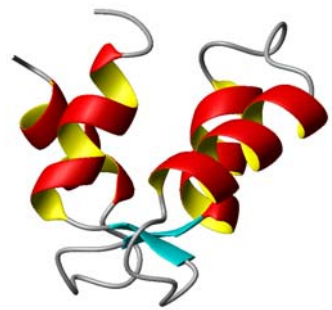


Target truncation

Target:
At3g03410.1







Jikui Song et al.