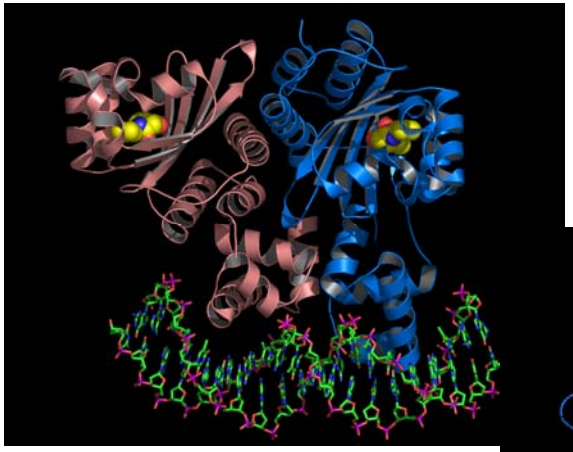
Structural Genomics
Exploring Protein Structure and Function

Andrzej Joachimiak

Structural Biology Center and
Midwest Center for Structural Genomics

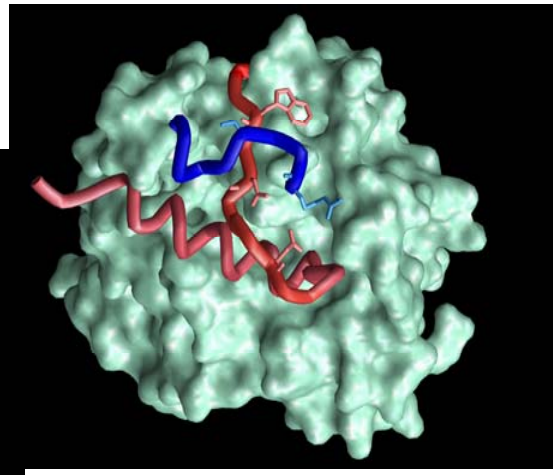
ORNL, July 12/13, 2005

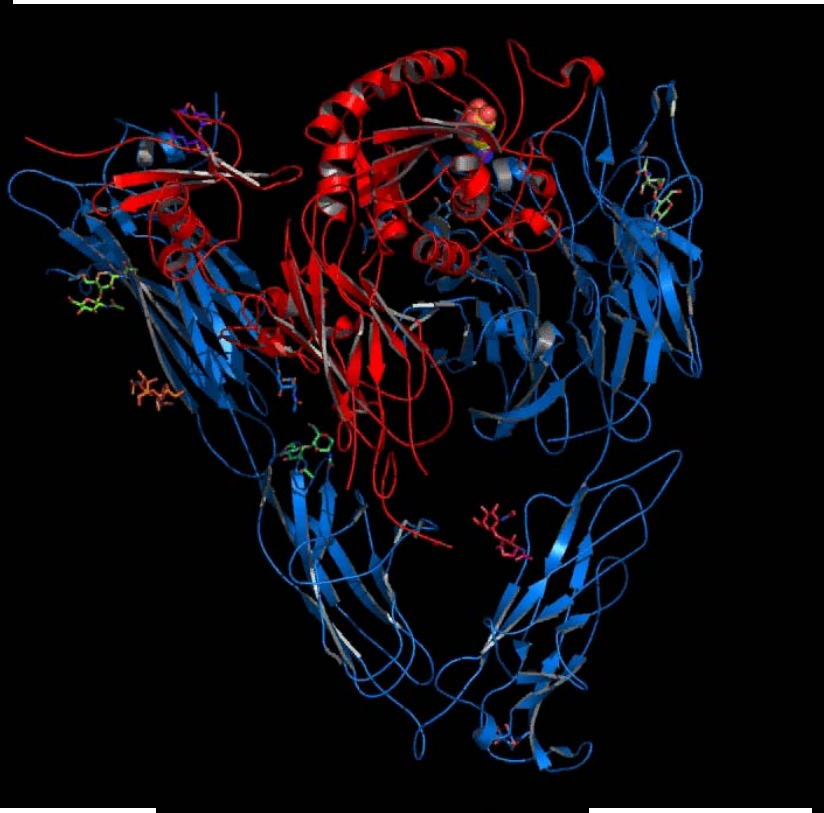
The Protein World



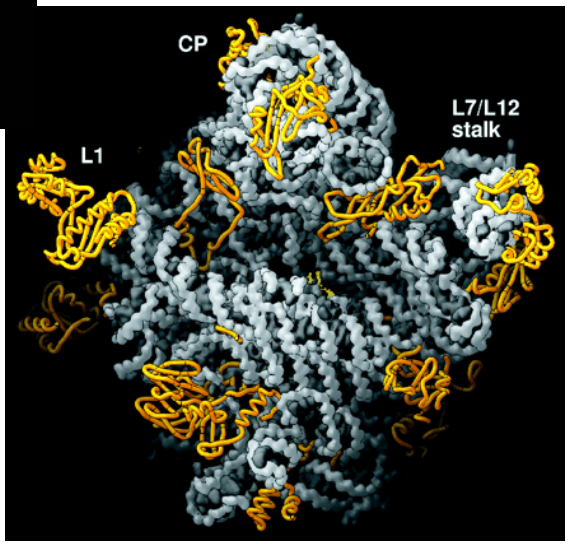
TraR, a pheromone-binding transcriptional factor

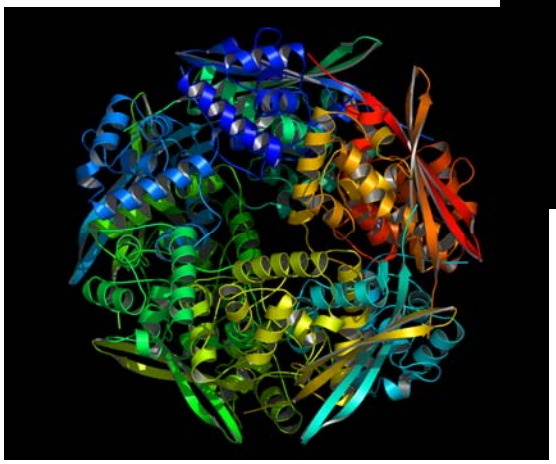
Protein with a knot



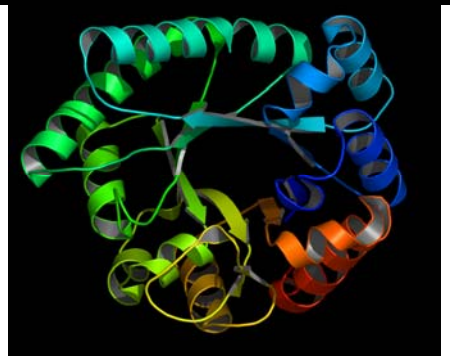


Ribosomal 50S subunit, multi-protein/RNA complex



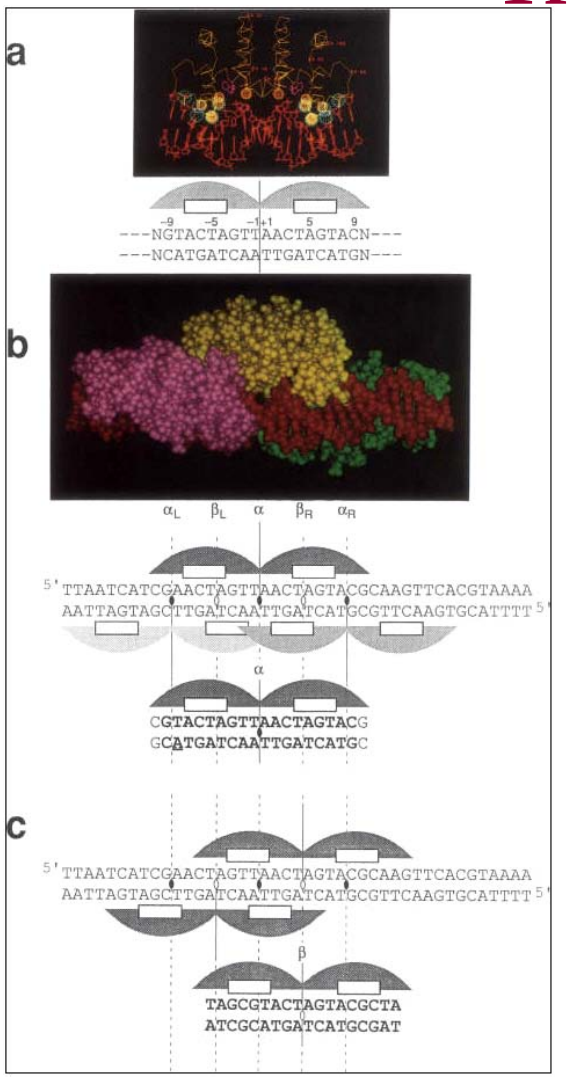


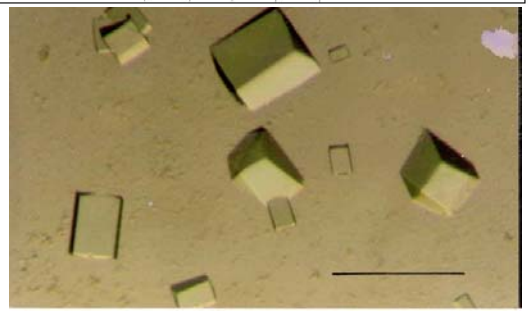
Cyanase, a homo-oligomeric enzyme

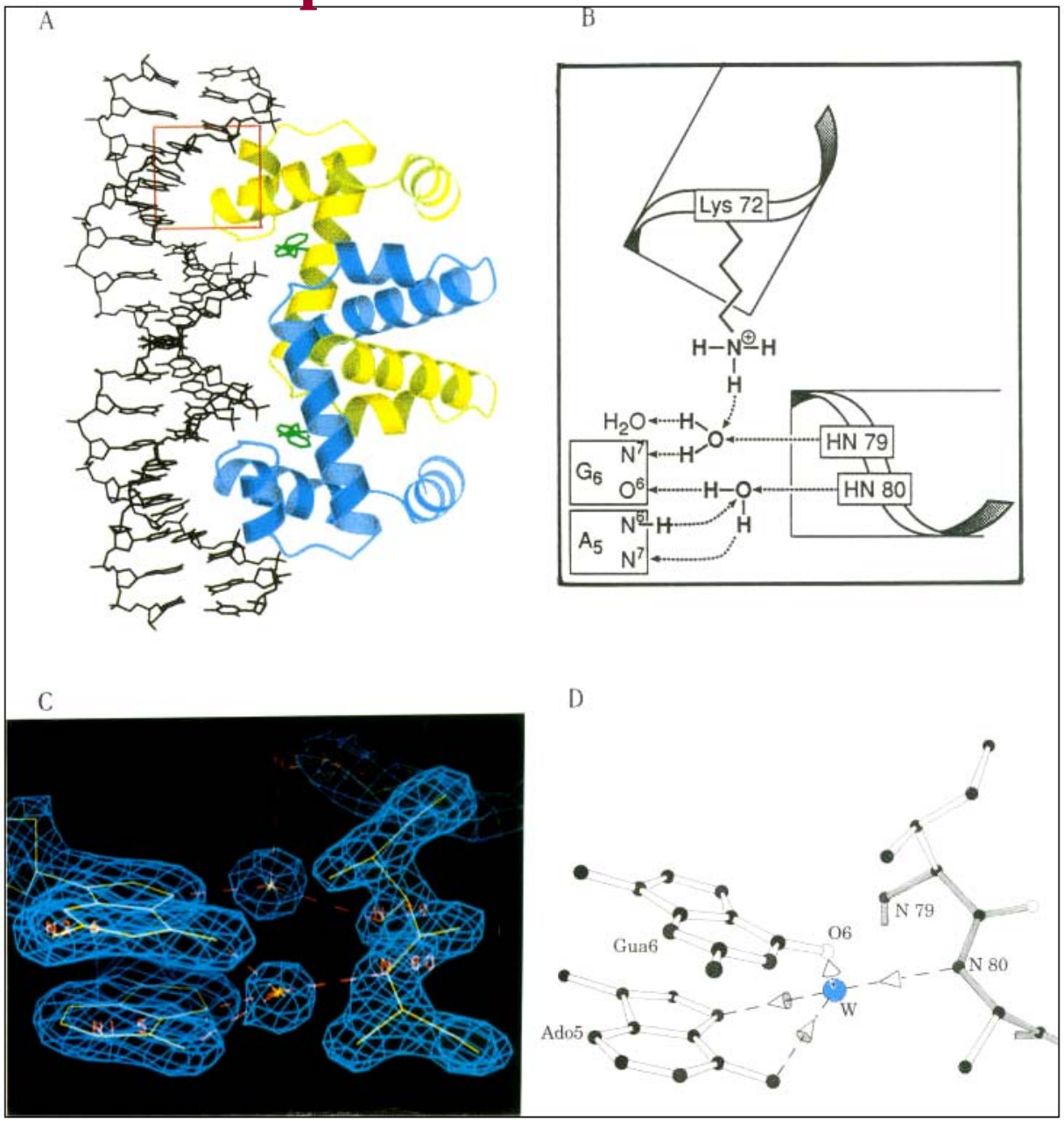


IolI, a monomeric enzyme

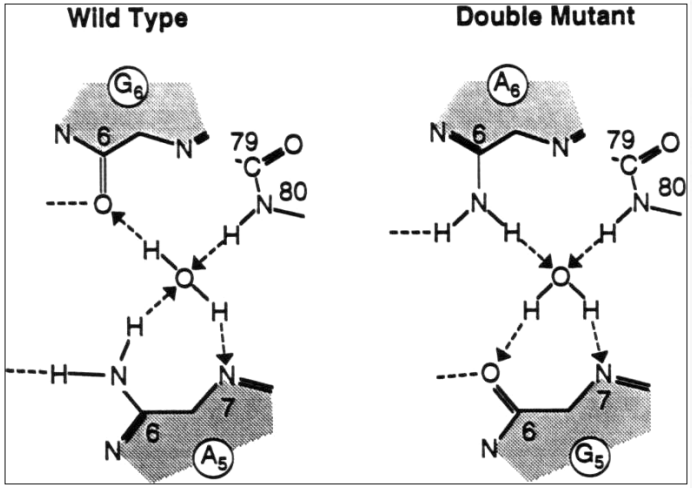
TrpR and its Complex with L-Trp and *trp* Operator DNA

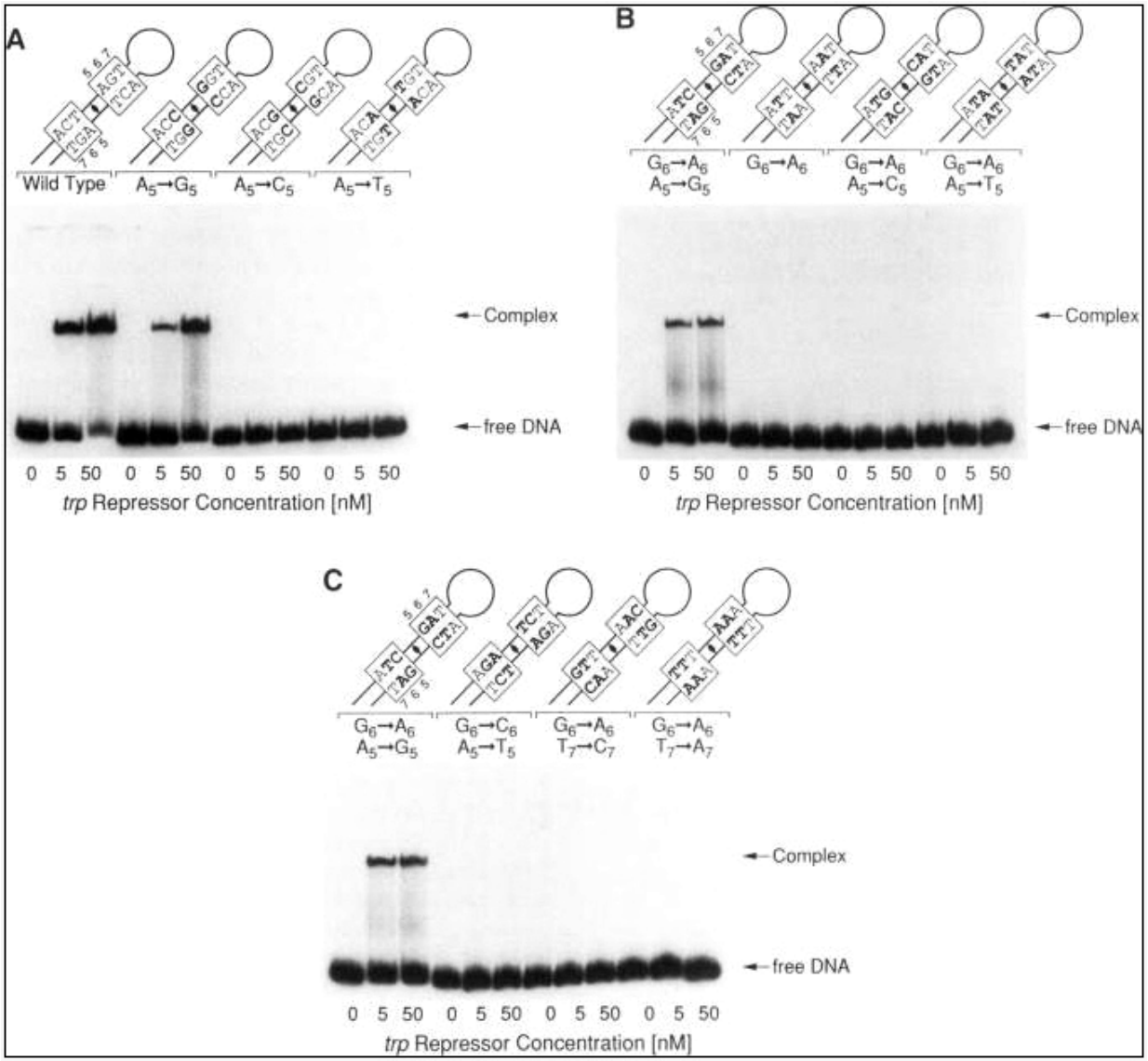




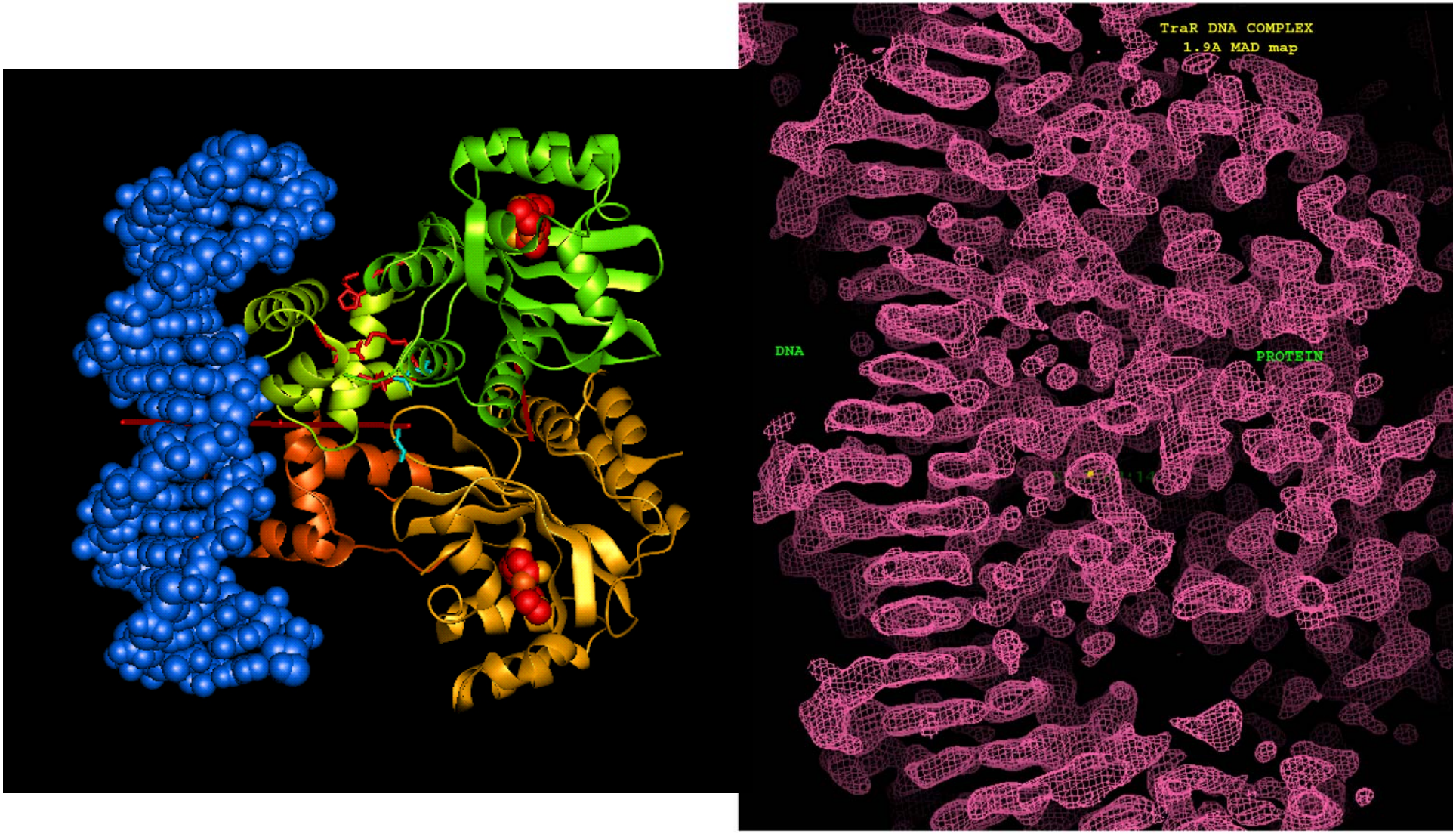


trp Operator Double Mutant Binds TrpR Specifically

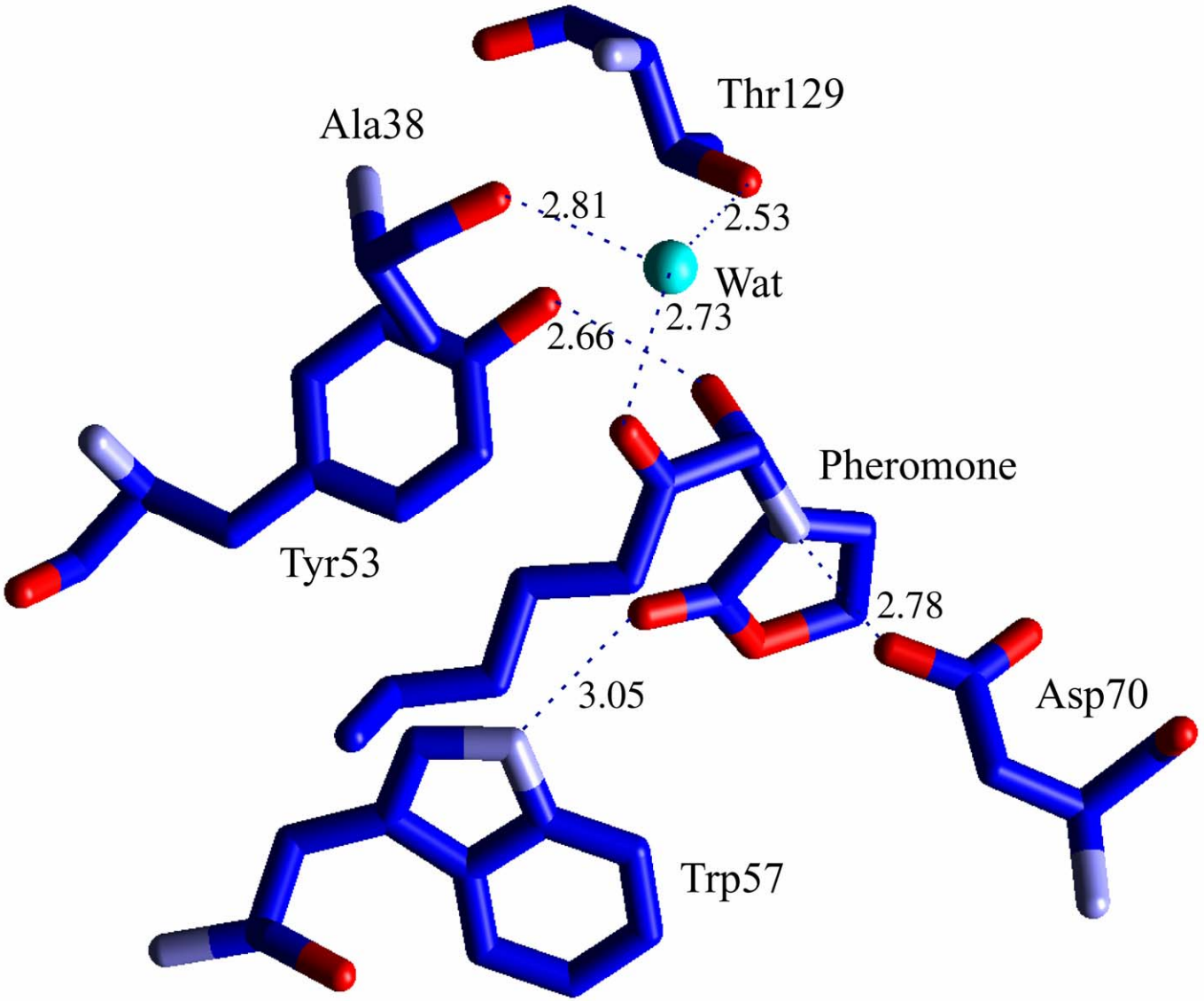




TraR/Pheromone/DNA Complex at 1.9 Å



Pheromone Binding Pocket – HSL is Fully Embedded within the Protein



Genomic Information is Being Accumulated Rapidly

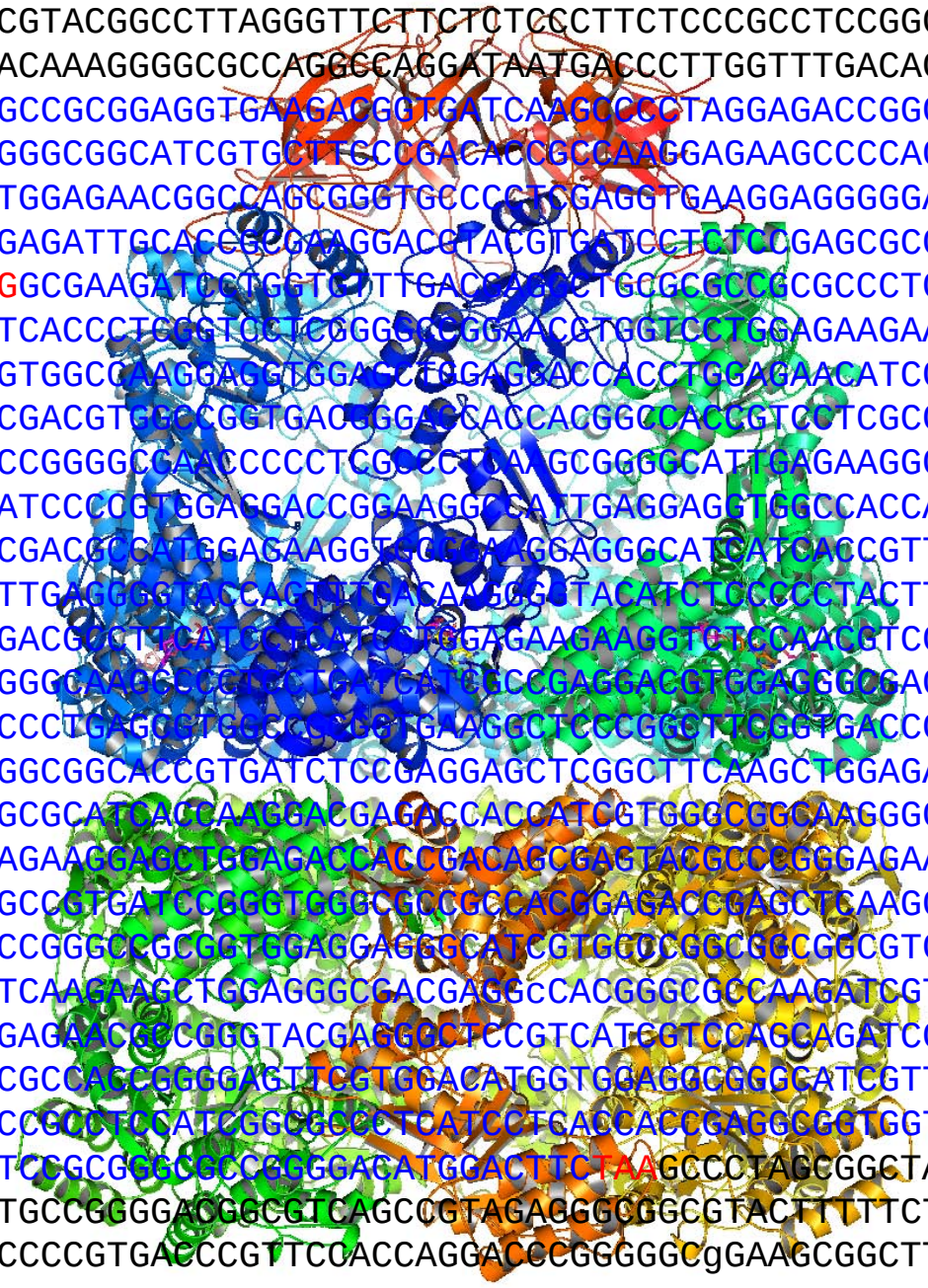
GOLD™
Genomes OnLine Database

Contact: Genomesonline	Last Update: March 22, 2005	Location www.genomesonline.org
	Search GOLD: 1421 genome projects	
261 Published Complete Genomes	669 Prokaryotic Ongoing Genomes	489 Eukaryotic Ongoing Genomes including 12 chromosomes

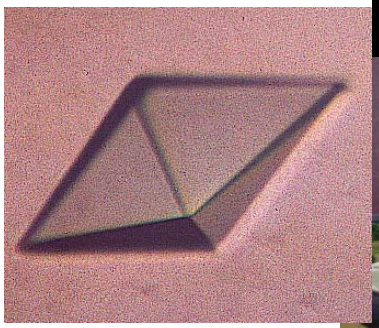
	ORFs	Seq. identity >30% PDB	In PDB
• *Salmonella typhimurium*	4597	1237 (26.9)	239
• *Shigella flexneri*	4068	1089 (26.8)	27
• *Yersinia pestis*	4090	1090 (26.7)	16
• *Haemophilus influenzae*	1709	608 (35.6)	128
• *Bacillus anthracis*	5311	1059 (20.0)	39
• *Bacillus cereus*	*5738*	1106 (19.3)	77
• *Listeria monocytogenes*	2846	788 (27.7)	18
• *Staphylococcus aureus*	2632	712 (27.0)	326
• *Vibrio cholerae*	2742	782 (28.5)	78

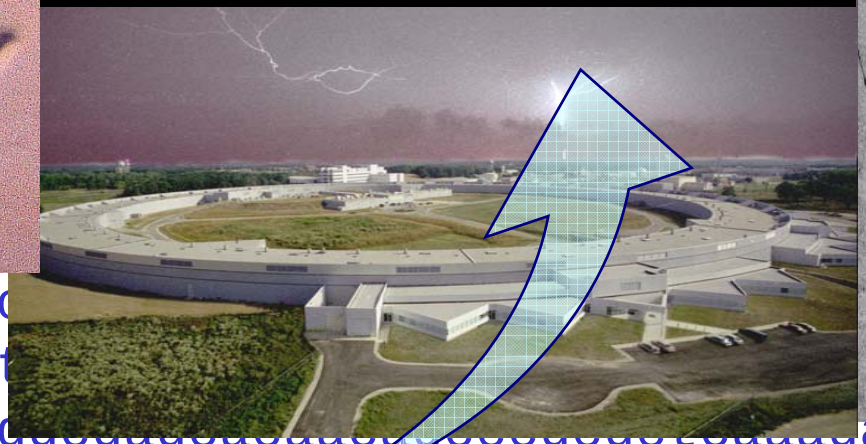
Hsp10/60 Operon Codes for Protein Folding Machine

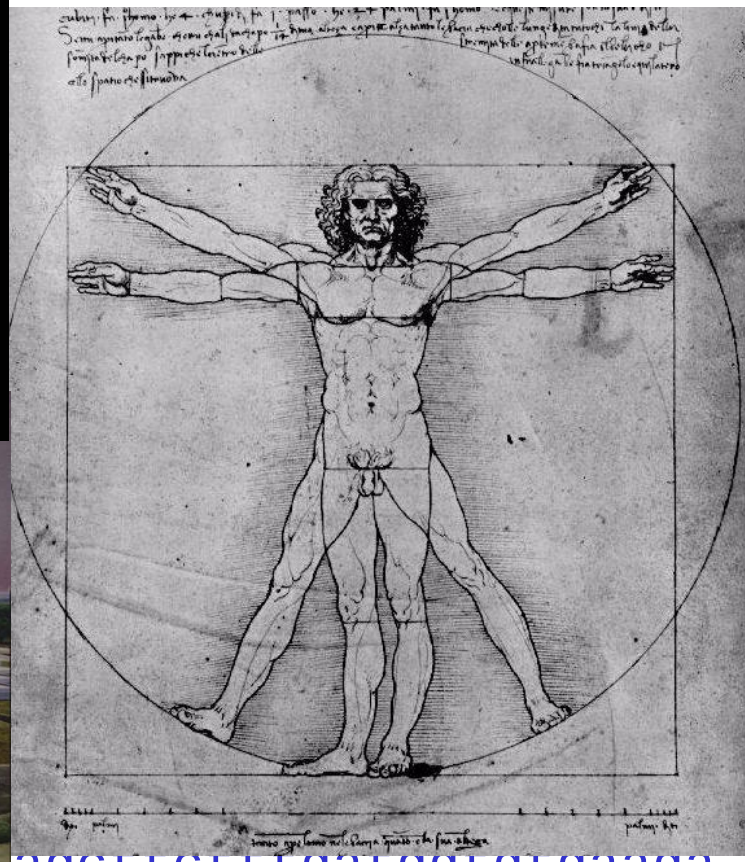
CGCTTTTGGAGCTCATCCGTACGGCCTTAGGGTTCTTCTCTCCCTTCTCCCGCCTCCGGGGTACAAGGCGATTTACTAGC
AAGCGCCAACCCCCTTGACAAAGGGGCGCCAGGCCAGGATAATGACCCTTGGTTTGACACTCTCACCCTTTGGGCGTCAA
AGGAGGGAGTGGATATGGCCGCGGAGGTGAAGACGGTGATCAAGCCCCTAGGAGACCGGGTGGTGGTGAAGCGCATTGAG
GAGGAGCCCAAGACCAAGGGCGGCATCGTGCTTCCCGACACCGCCAAGGAGAAGCCCCAGAAGGGCAAGGTGATCGCGGT
GGGCACGGGCCGCGTCTTGGAGAACGGCCAGCGGGTGCCCCTCGAGGTGAAGGAGGGGGACATCGTGGTCTTCGCCAAGT
ACGGCGGCACCGAGATTGAGATTGCACCGCGAAGGACGTACGTGATCCTCTCCGAGCGCGACCTGCTTGCGGTCCTGCAG
TAAAGGAGGTGAACCATGGCGAAGATCCTGGTGTTTGACGAGGCTGCGCGCCGCGCCCTGGAGCGCGGGGTCAACGCGGT
GGCCAACGCGGTGAAGGTCACCCTCGGTCCTCGGGGCGGAACGTGGTCCTGGAGAAGAAGTTTGGCTCCCCCACCATCA
CCAAGGACGGGGTGACGGTGGCCAAGGAGGTGGAGCTGGAGGACCACCTGGAGAACATCGGGGCCCAGCTCCTCAAGGAG
GTGGCCTCCAAGACCAACGACGTGGCCGGTGACGGGACCACCACGGCCACCGTCCTCGCCCAGGCCATCGTCCGGGAGGG
CCTGAAGAACGTGGCCGCCGGGGCCAACCCCCTCGGCCTCAAGCGGGGCATTGAGAAGGCGGTGGAGGCCGCGGTGGAGA
AGATCAAGGCCCTCGCCATCCCCGTGGAGGACCGGAAGGCCATTGAGGAGGTGGCCACCATCTCCGCCAACGATCCCGAG
GTCGGCAAGCTGATTGCCGACGCCATGGAGAAGGTGGGCAAGGAGGGCATCATCACCGTTGAGGAGTCCAAGAGCCTCGA
GACCGAGCTCAAGTTCGTTGAGGGGTACCAGTTTGACAAGGGGTACATCTCCCCCTACTTCGTCACCAACCCCGAGACGA
TGGAGGCGGTCCTCGAGGACGCCTTCATCCTCCTCGTGGAGAAGAAGGTGTCCAACGTCCGTGAGCTCCTCCCCATCCTG
GAGCAGGTGCGCCAGACGGGCAAGCCCCTGCTGATCATCGCCGAGGACGTGGAGGGCGAGGCTTTGGCCACCCTGGTGGT
GAACAAGCTCCGGGGCACCCTGAGGGTGGCCGCGGTGAAGGCTCCCGGCTTCGGTGACCGCAGGAAGGAGATGCTCAAGG
ACATCGCGGCCGTCACCGGCGGCACCGTGATCTCCGAGGAGCTCGGCTTCAAGCTGGAGAACGCCACCCTCTCCATGCTG
GGCCGGGCCGAGCGGGTGCGCATCACCAAGGACGAGACCACCATCGTGGGCGGCAAGGGCAAGAAGGAGGACATTGAGGC
CCGGATCAACGGCATCAAGAAGGAGCTGGAGACCACCGACAGCGAGTACGCCCGGGAGAAGCTCCAGGAGCGCCTGGCGA
AGCTCGCGGGTGGCGTGGCCGTGATCCGGGTGGGCGCCGCCACGGAGACCGAGCTCAAGGAGAAGAAGCACCGCTTTGAG
GACGCCCTGAACGCCACCCGGGCCGCGGTGGAGGAGGGCATCGTGCCCGGCGGCGGCGTGACCCTCCTCCGGGCCATCAG
CGCCGTGGAGGAGCTCATCAAGAAGCTGGAGGGCGACGAGGCCACGGGCGCCAAGATCGTGCGCCGGGCCCTGGAGGAAC
CCGCCCGCCAGATCGCCGAGAACGCCGGGTACGAGGGCTCCGTCATCGTCCAGCAGATCCTGGCCGAGACCAAGAACCCC
CGCTACGGCTTCAACGCCGCCAGCGGGGAGTTCGTGGACATGGTGGAGGCGGGGATCGTTGACCCCGCCAAGGTGACCCG
CTCtGCCCTGCAGAACGCCGCCTCCATCGGCGCCCTCATCCTCACCACCGAGGCGGTGGTGGCGGAGAAGCCCGAGAAGA
AGGAGTCCACCCCCGCCTCCGCGGCGCCGGGGACATGGACTTCTAAGCCCTAGCGGCTAAGGGGCCCGGGCCTTCAAGG
CCCGGCCCCTTTCTAAGTGCCGGGGACGGCGTCAGCCGTAGAGGGCGGCGTACTTTTTCTCCAGGTAGGCCAGGAAGGGC
CTGGCGCTCGGGGCCTCCCCCGTGACCCGTTCCACCAGGACCCGGGGGCgGAAGCGGCTTCCCTCGGCGTGGATCC



From Gene to Structure and Function







Structural Genomics and Protein Structure Initiative

- **Exploring Protein Folding Space**
  - **Experimental determination of novel proteins using X-ray crystallography or NMR**
  - **Computer homology modeling of protein structures using protein folds database**
- **Approach**
  - **Cluster protein sequences (ORFs) in all known genomes into homologous groups (superfamilies and families)**
  - **Determine structure of few members of each family**
  - **Computer model all other members of family**
- **Using this approach it is believed that we need to determine structures of 15,000 – 20,000 members of protein families to cover 90% of protein fold space**

Midwest Center for Structural Genomics

PSI Protein Structure Initiative

www.mcsg.anl.gov

W. Anderson, A. Edwards, D. Fremont, A. Joachimiak,
W. Minor, Z. Otwinowski, C. Orengo, J. Thornton,

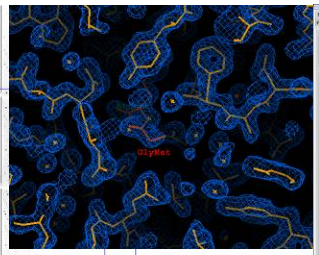
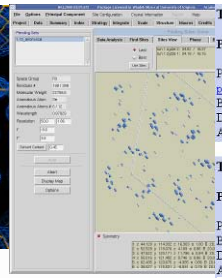
How Many Targets Are There?

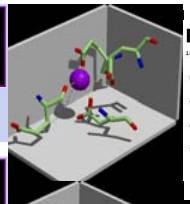
- **A novel protein family landscape protocol was developed at UCL by C. Orengo and allowed clustering of ORFs:**
  - **>1,000,000 sequences from 150 completed genomes.**
  - **These genes were clustered into >50,000 protein families.**
  - **There were >150,000 singleton sequences unique to specific organisms (10-15%/genome).**
  - **>14,500 genes in these genomes were labelled as 'priority 1' targets and share no detectable sequence similarity to a sequence of structure in the PDB.**
  - **6,190 Pfam families, 2,108 with homologue in PDB.**
  - **1,200 very large Pfam families with no structural homologue cover 70% of sequence space.**
- **Currently there are ~2,500,000 ORFs in public databases and ~27,000 protein structures (~4,300 nonredundant)**

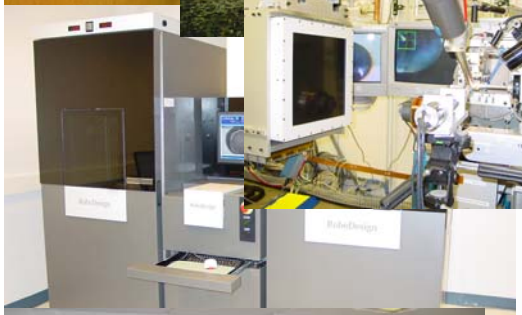
Midwest Center for Structural Genomics

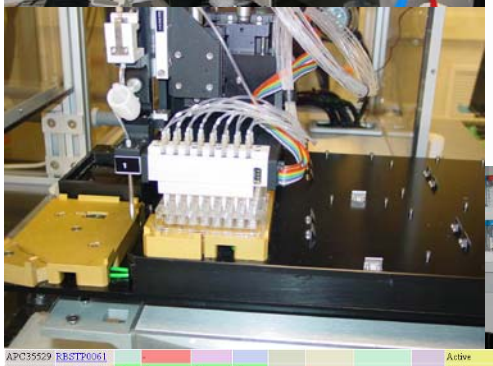
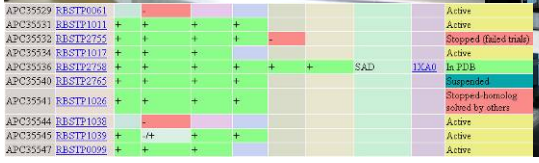
PSI Protein Structure Initiative



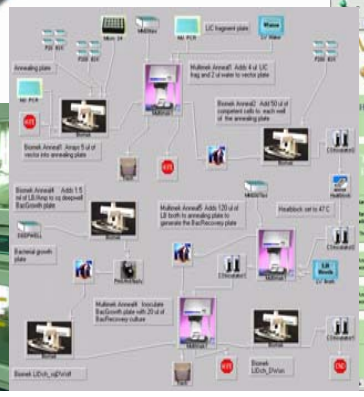
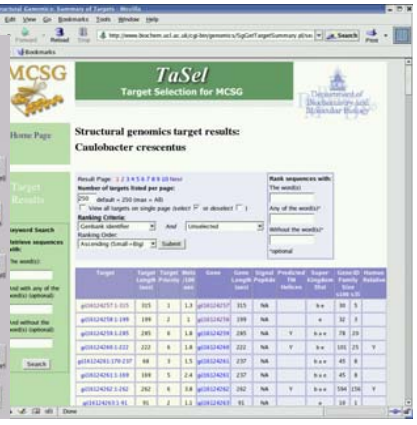


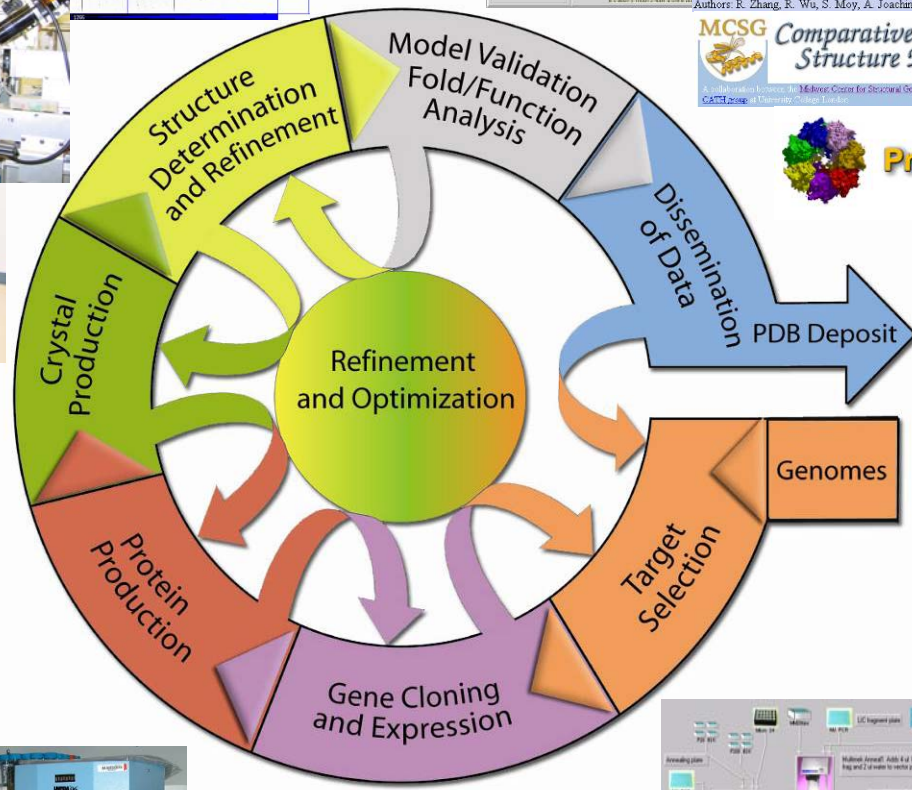




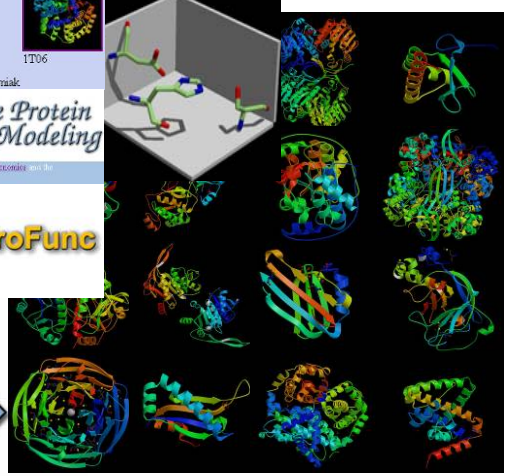




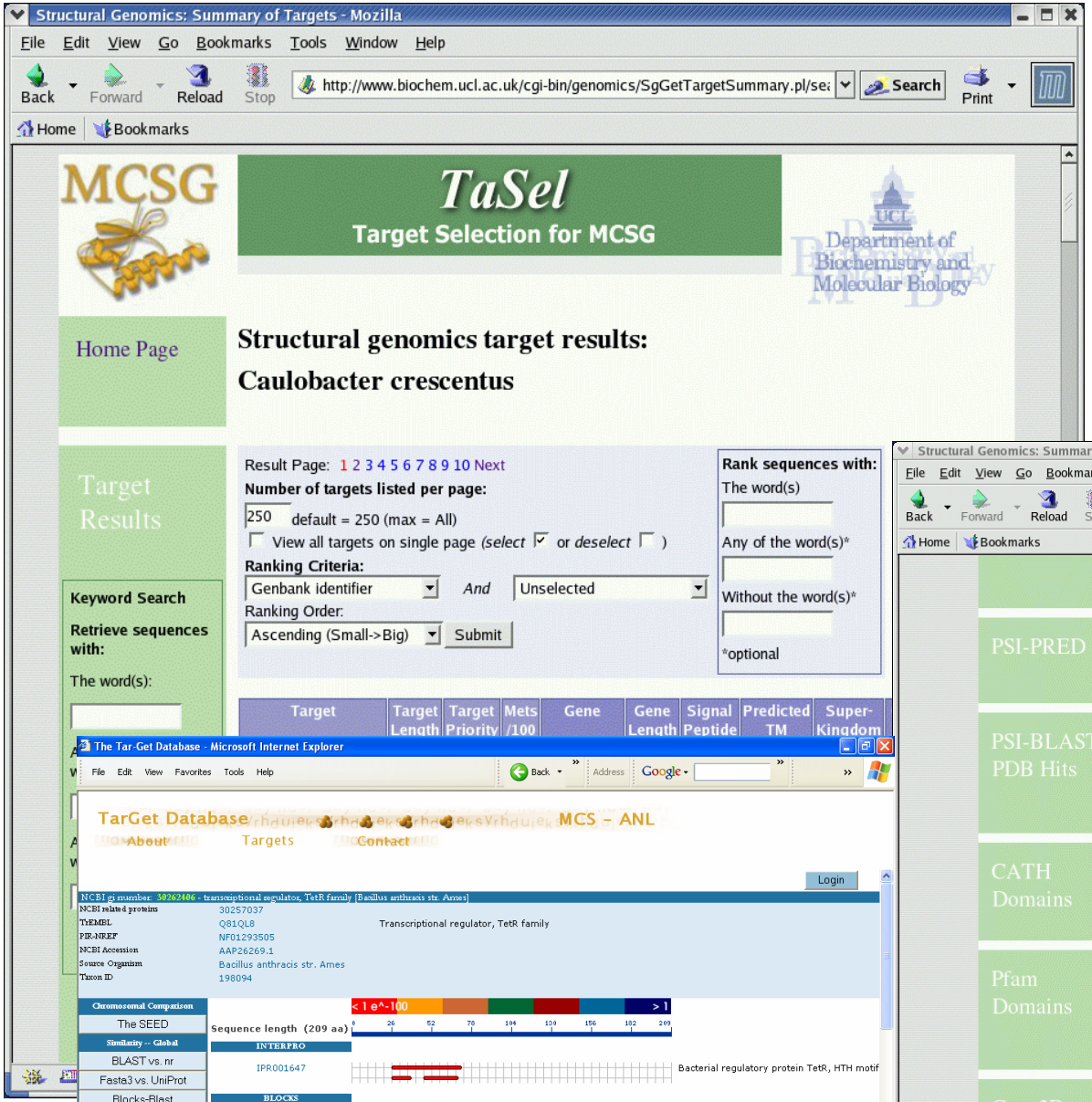
MCSG Comparative Protein Structure Modeling

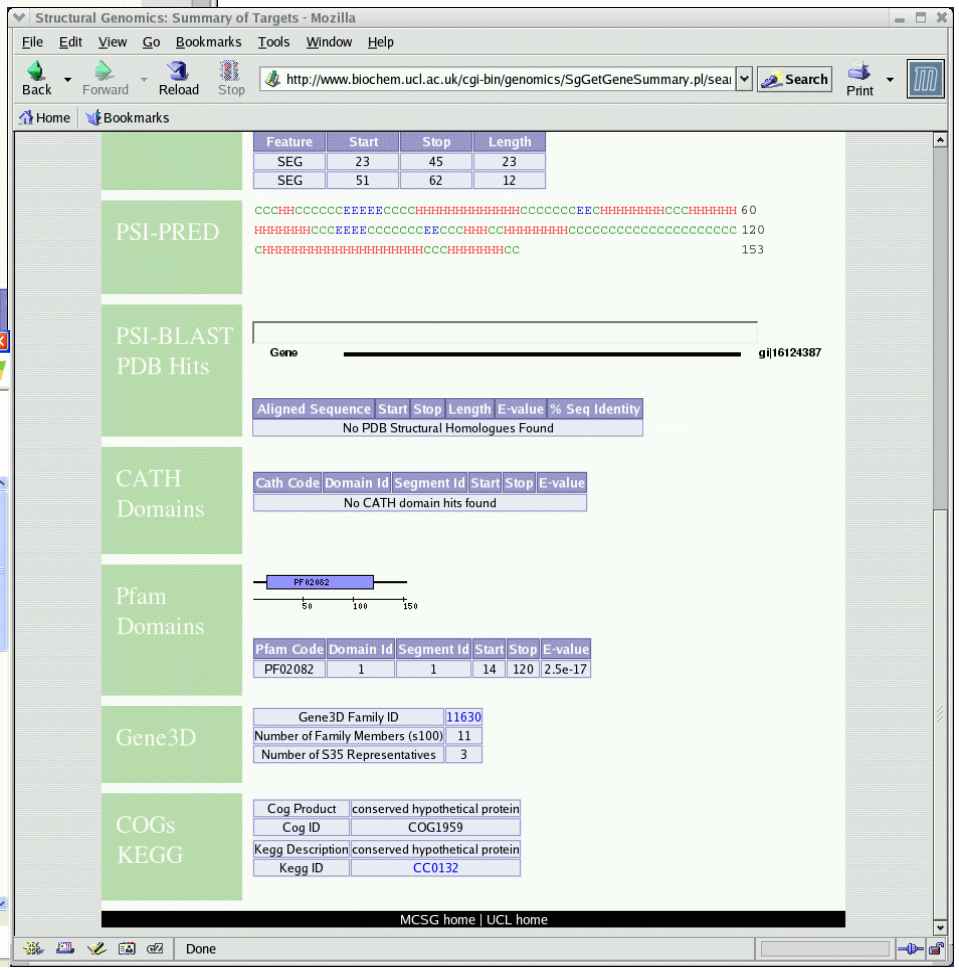
ProFunc



ATGAGAATGAAGCGATTTTTAA
TTTGGTTACAAAATTGTTCAAA
AAATTACAACAAAAATTTCAAA
TTTGAATCACTTGAAAAAATAA
TCATTAAATTATCCAGTACTAC
GAGCGAGAACATCGACGTAAAG

Target Selection Databases: TaSel, Gene3D, and TargetDB





Gene Cloning and Protein Expression Pipeline

Express Primer Tool

Fragment Amplification

Picogreen Analysis

LIC Fragment Preparation

Annealing Reaction

Plasmid Isolation

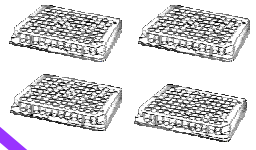
Immuno Analysis

Expression Analysis

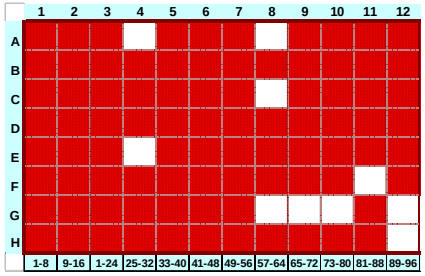
Solubility Analysis

Distribution and Storage

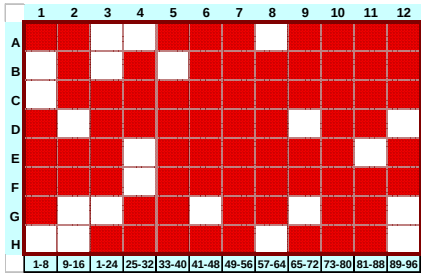
Plasmid plates



Amplification

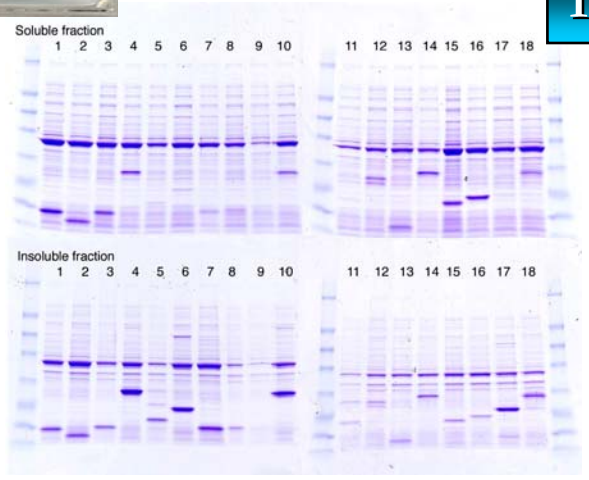


Expression

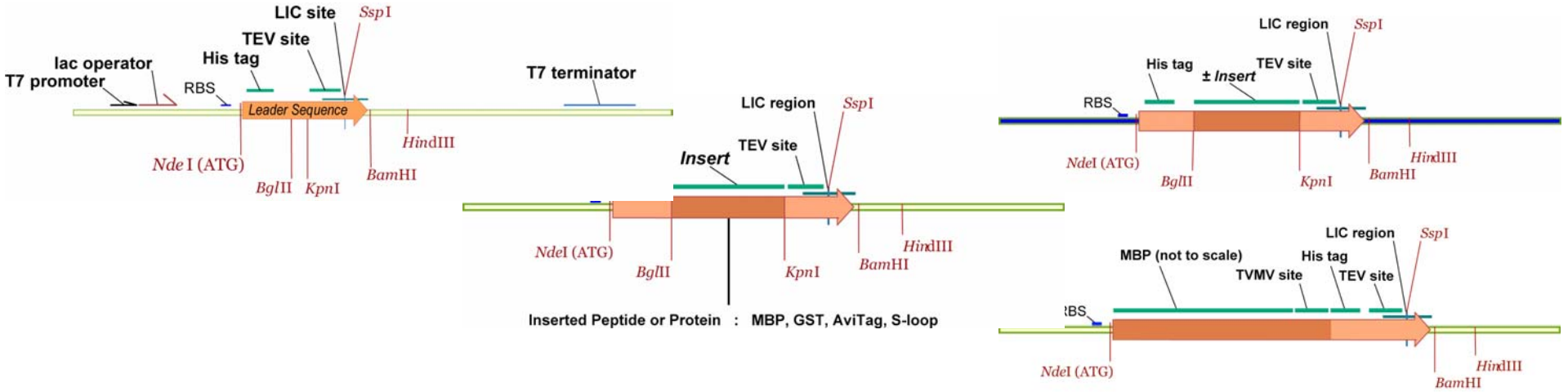


Cryovials

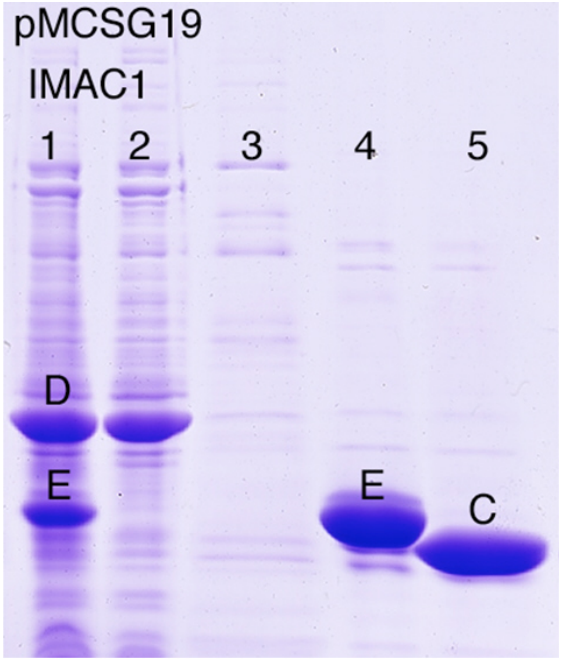
Solubility



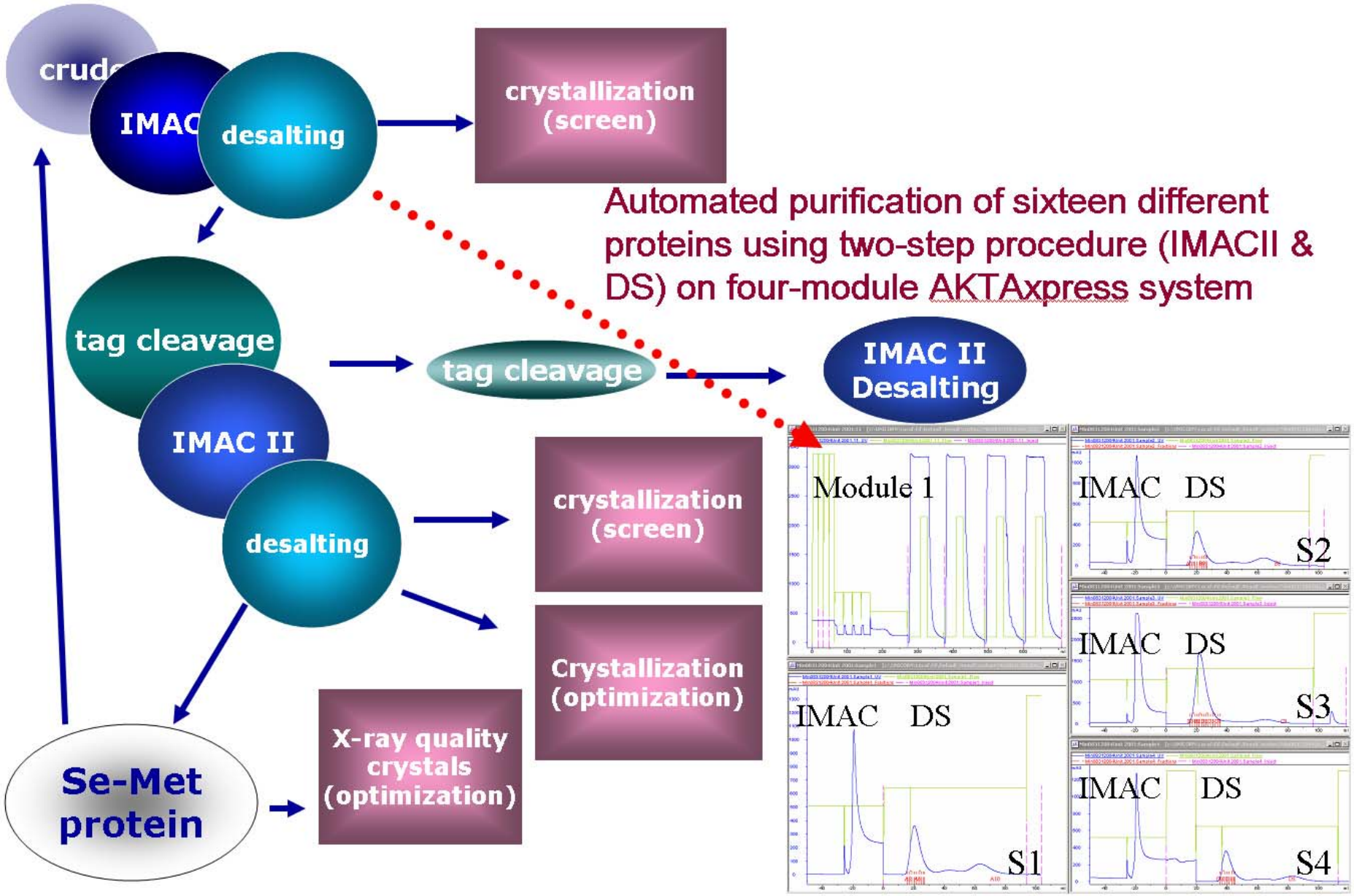
Vectors Constructed for HTP Production of Proteins



Vector	Base Vector	Encoded Leader Sequence	Use
pMCSG7	pET21a	N-His-TEV-LICs-	Routine protein production
pMCSG8	pMCSG7	N-His-Sloop-TEV-LICs	Improve solubility
pMCSG9	pMCSG7	N-His-MBP-TEV-LICs	Improve solubility
pMCSG10	pMCSG7	N-His-GST-TEV-LICs	Improve solubility
pMCSG11	pACYCDuet-1	N-His-TEV-LICs	Coexpression
pMCSG12	pACYCDuet-1	N-His-Sloop-TEV-LICs	Coexpression
pMCSG13	pACYCDuet-1	N-His-MBP-TEV-LICs	Coexpression
pMCSG14	pACYCDuet-1	N-His-GST-TEV-LICs	Coexpression
pMCSG17	pMCSG7	N-Stag-TEV-LICs	Coexpression
pMCSG20	pMCSG7	N-Stag-GST-TEV-LICs	Coexpression
pMCSG16	pMCSG7	N-His-AviTag-TEV-LICs	Phage display
pMCSG15	pMCSG7	LICs-TEV-AviTag-His-C	Phage display
pMCSG18	pMCSG7	N-His-TEV-LICs-GFP	Screening
PMCSG19	pMCSG7	N-MBP-TVMV-His-TEV-LICs	Purification
pMCSG21	pDONR/zeo	attL1-TEV-LIC-attL2	Gateway cloning



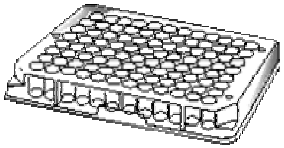
Automated Protein Purification for Structural Genomics

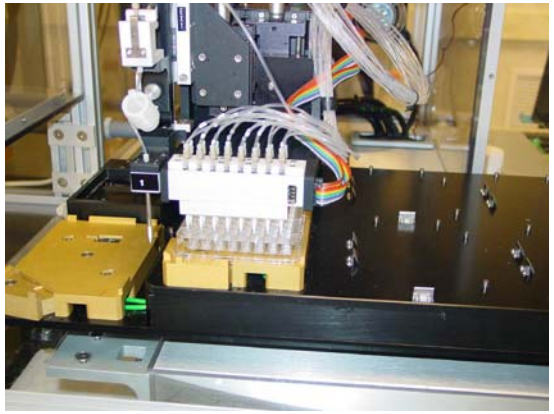


HTP Protein Crystallization

Commercial screens

Matrix Maker
Custom screens making

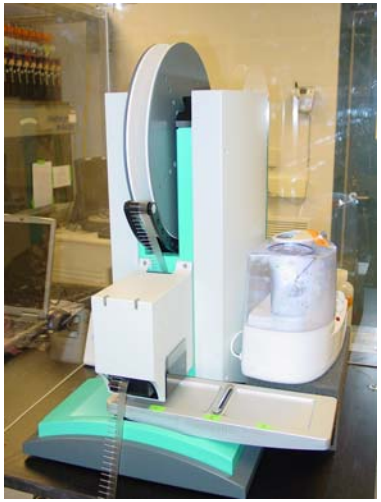




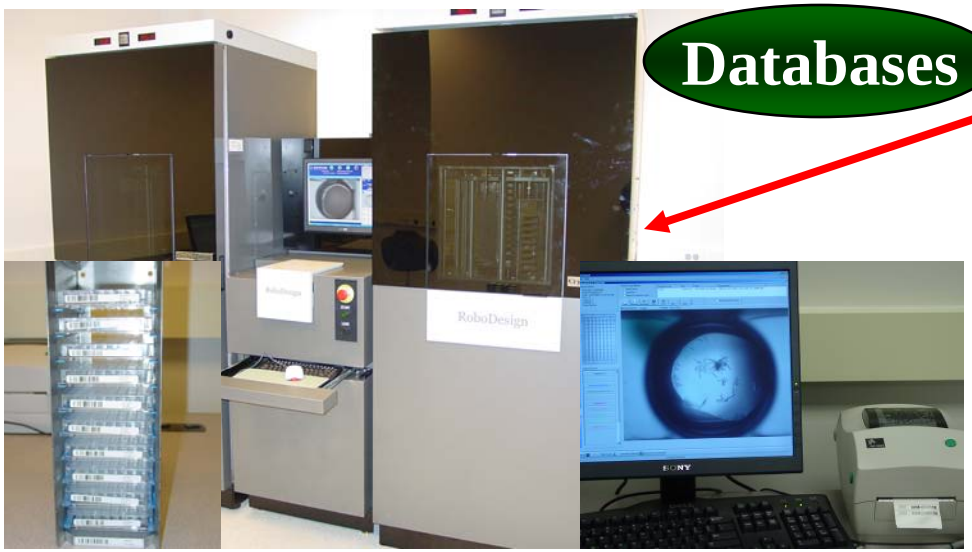
Hydra
Dispensing of reservoirs

Cartesian
Set up crystallization

Mosquito
Set up crystallization



Manual
Crystallizations

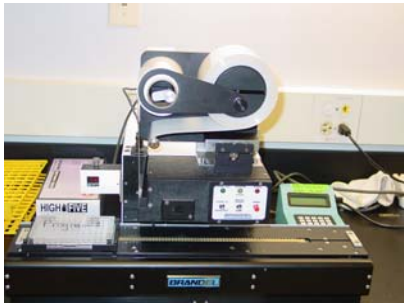


Databases

Visualization
Manual and robotic

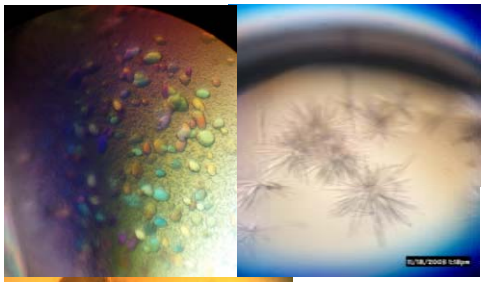
Crystal freezing

X-ray test & Data collection

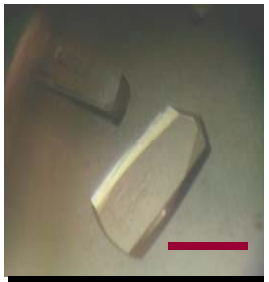


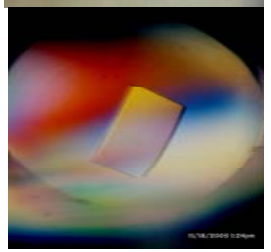
Brandel
Sealing boxes

Nanoliter Crystallization in 96-well Format



- **Increases crystallization throughput,**
- **Small volume reduces protein quantities needed for complete screening,**
- **Can be used with custom optimization screens,**





INDEX 96

PEG/ION 48

SaltRx 96

Wizard1 48

Wizard2 48

WizardCryo1 48

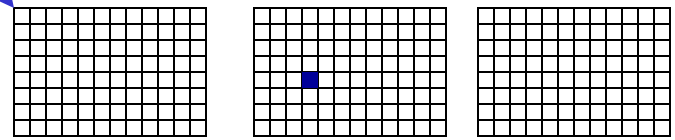
WizardCryo2 48

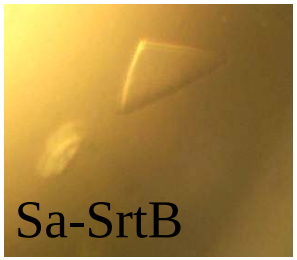
Hampton2 48

HamotonCryo 50

Hampton1 50

Screening throughput: ~20,000 droplets/8 hrs



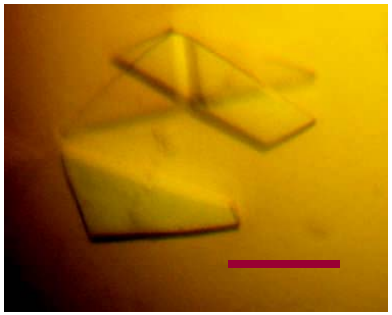


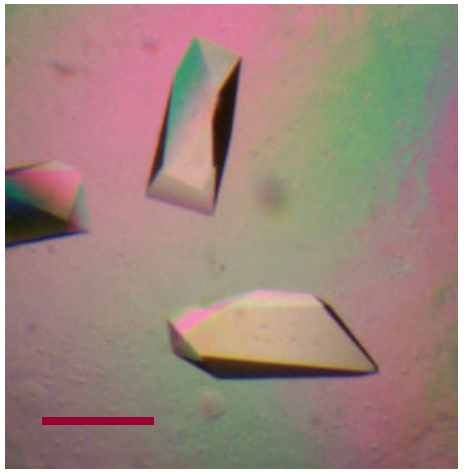
Format: 96-well conditions x 3 proteins

Optimization

Exotic | **PEG-based** | **Salt-based**

Fine tuning

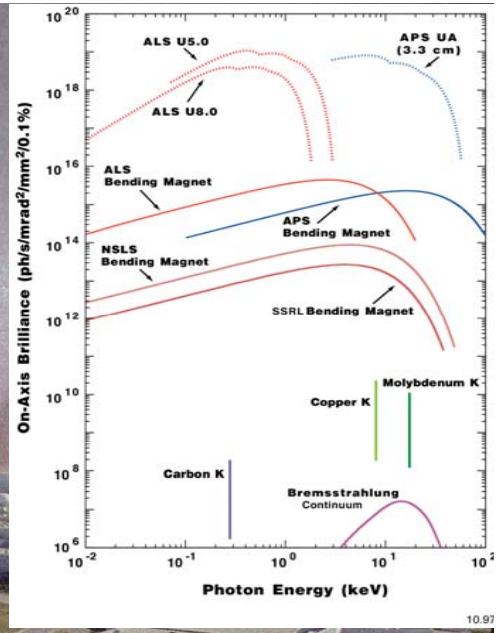




Third Generation Synchrotron Beamlines for Macromolecular Crystallography

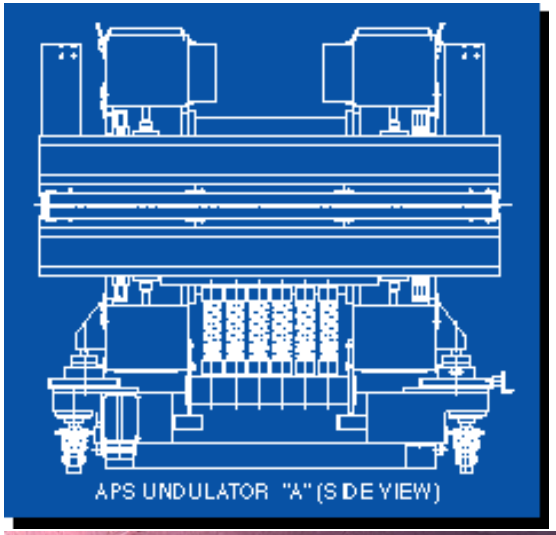
Advanced Photon Source:

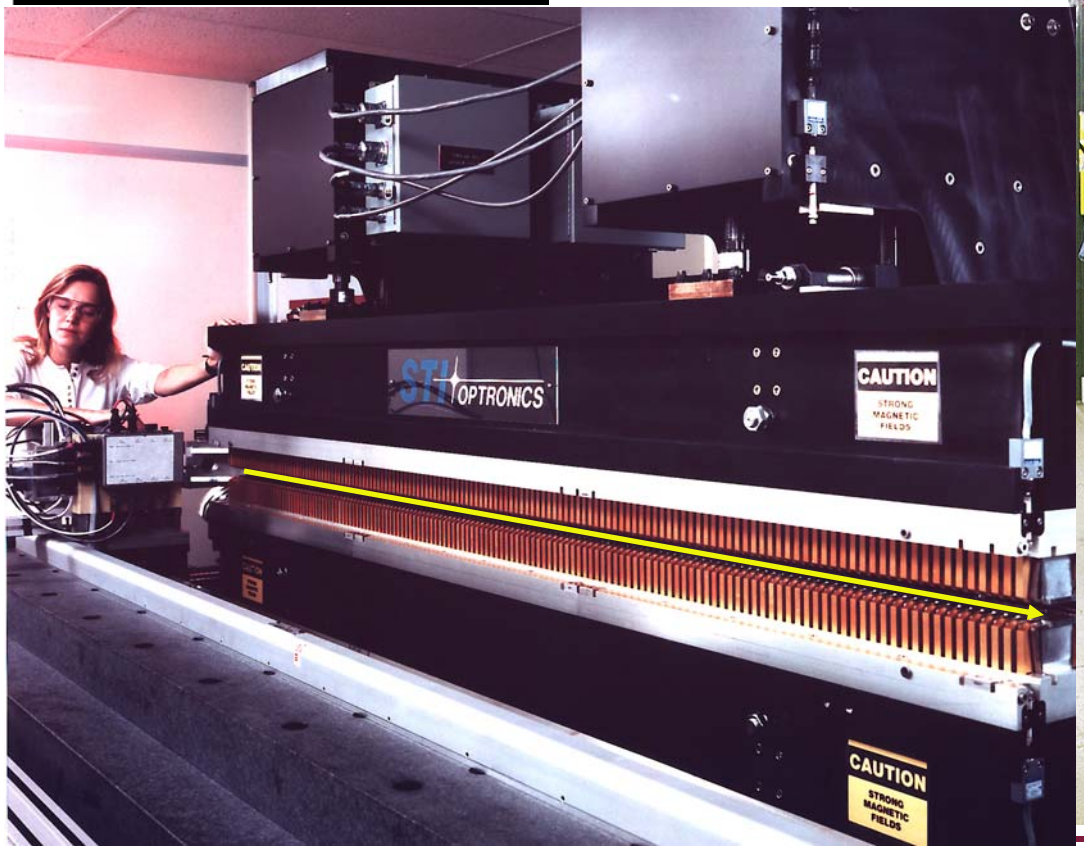
- Dedicated storage ring with enhanced spectral properties
- Intense, brilliant beams tailored to wide range of biological experiments
- Undulator source
- Anomalous diffraction
- Atomic resolution
- CCD area detectors
- Cryo-crystallography
- Advanced computing
- Optimized for speed and efficiency

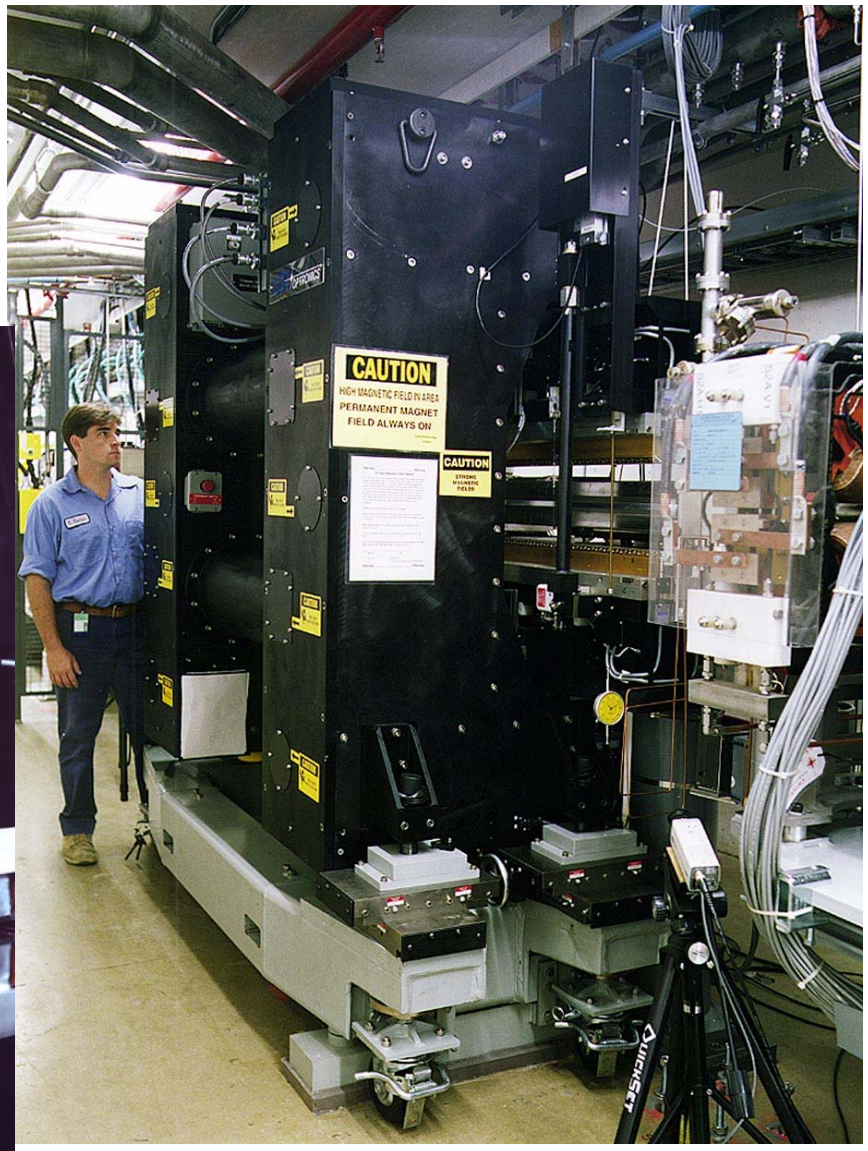


SBC

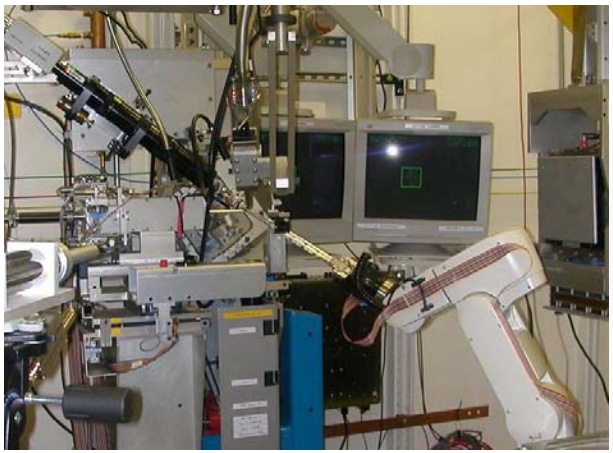
APS Undulator A



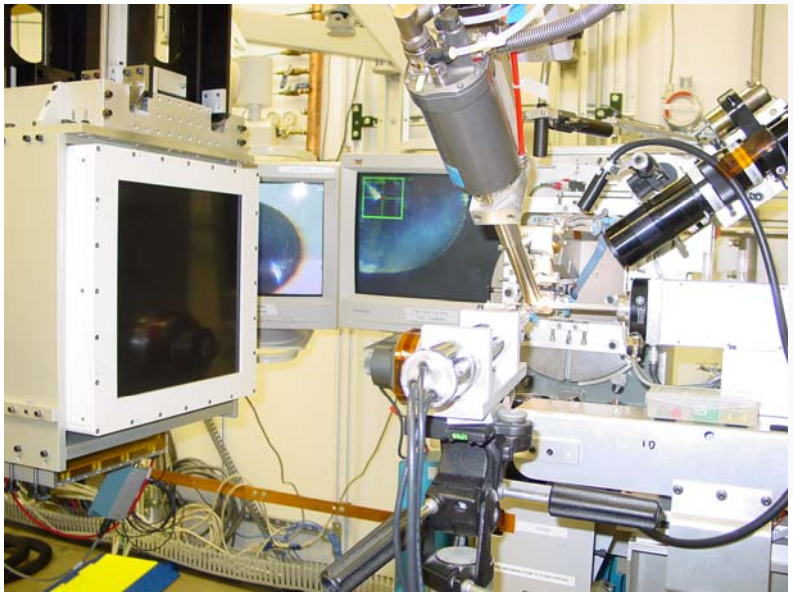


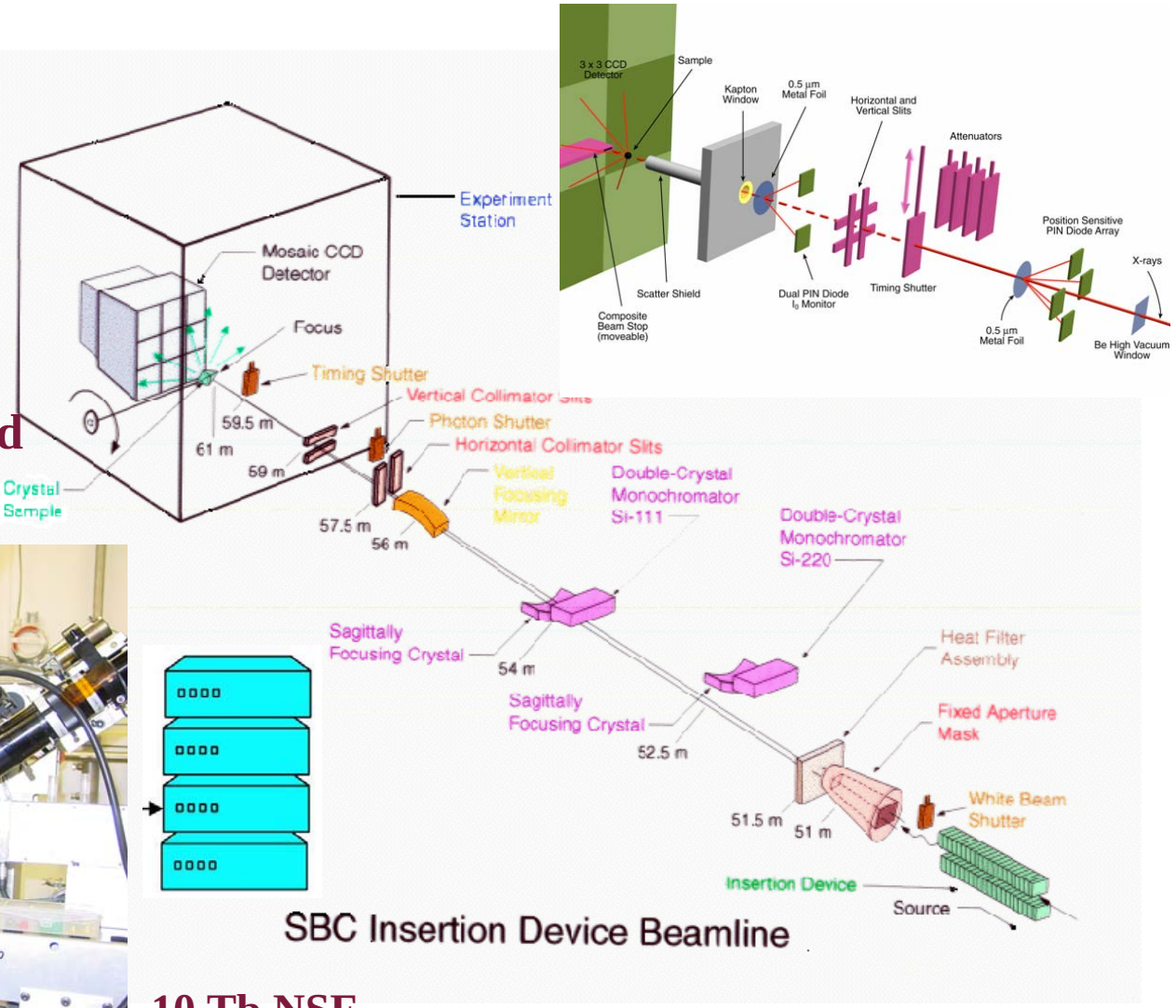


Dedicated X-ray Beamlines for Macromolecular Crystallography - SBC 19ID and 19BM Beamlines at the Advanced Photon Source – 851 PDB Deposits



Goniostat, Q315 detector and robot for crystal mounting



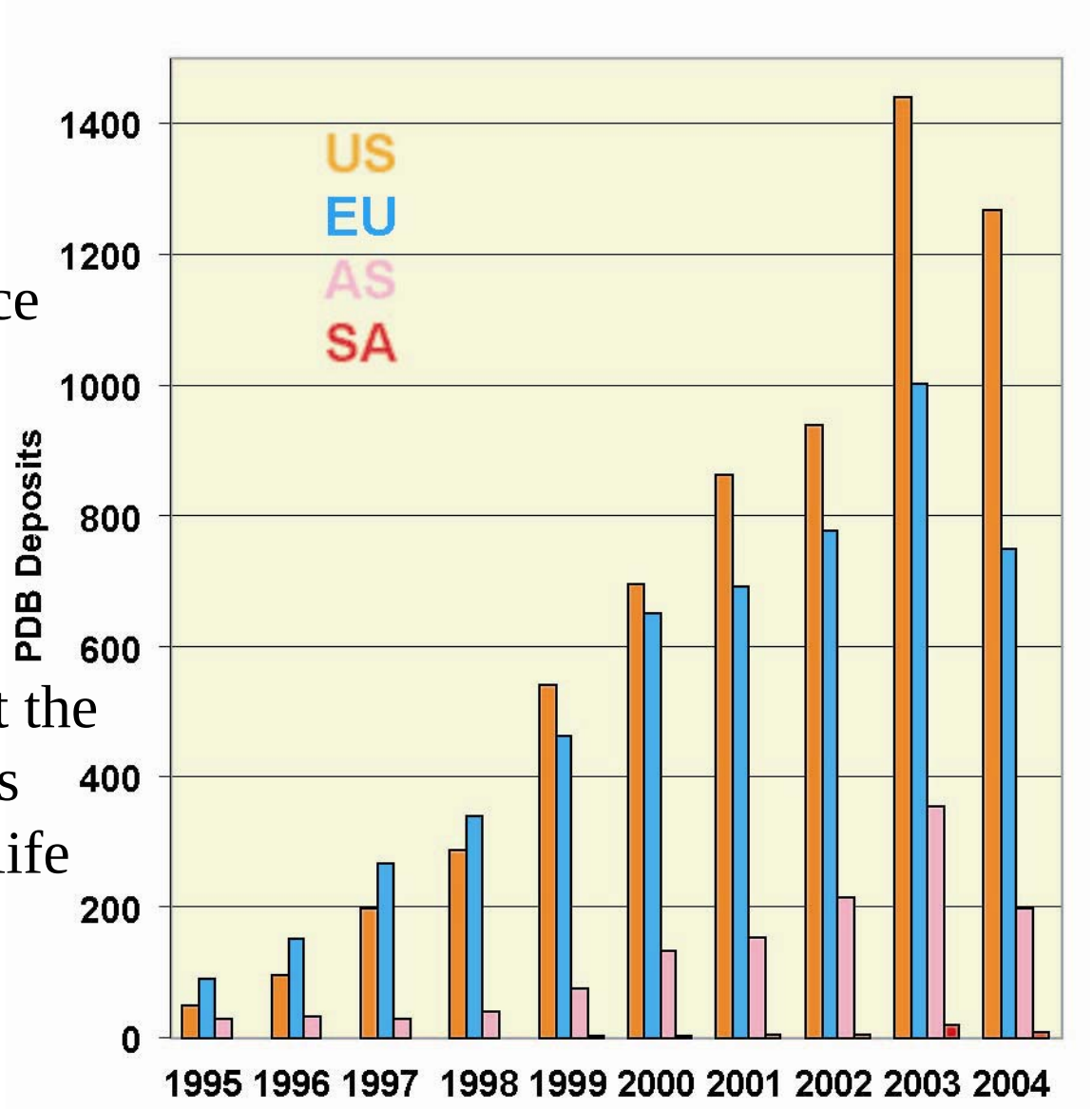


10 Tb NSF Storage

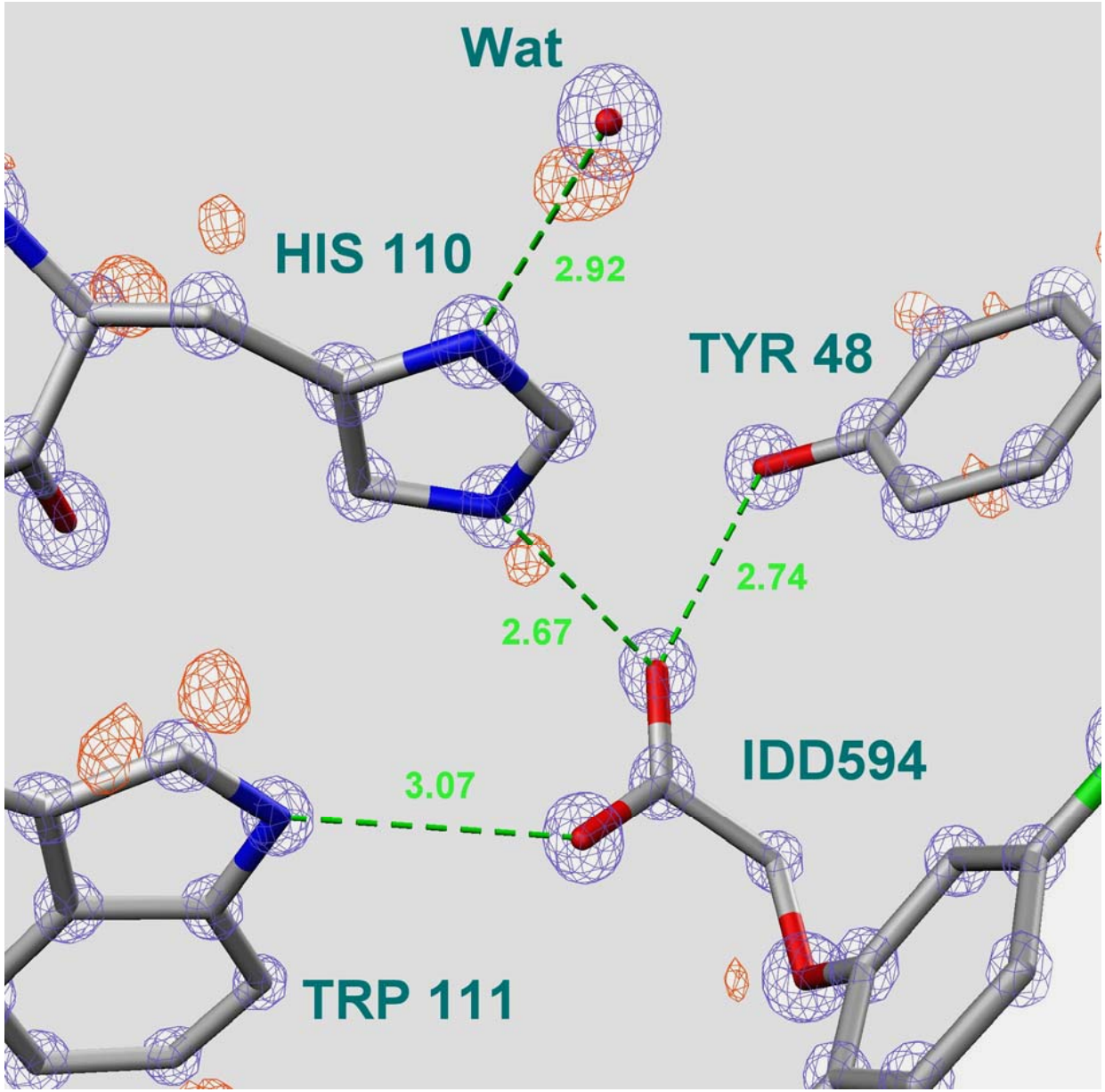
Structures of Proteins in PDB

- Single crystal x-ray crystallography ~23,000
- Nuclear magnetic resonance ~4,000
- Single crystals neutron crystallography ~few

- For x-ray data collection at the synchrotron cryofreezing is essential - extends crystal life facilitating better data



Aldose Reductase at 0.66 Å – Protonation of His110 – a Key Catalytic Residue



Blue Contours:

2Fo-Fc

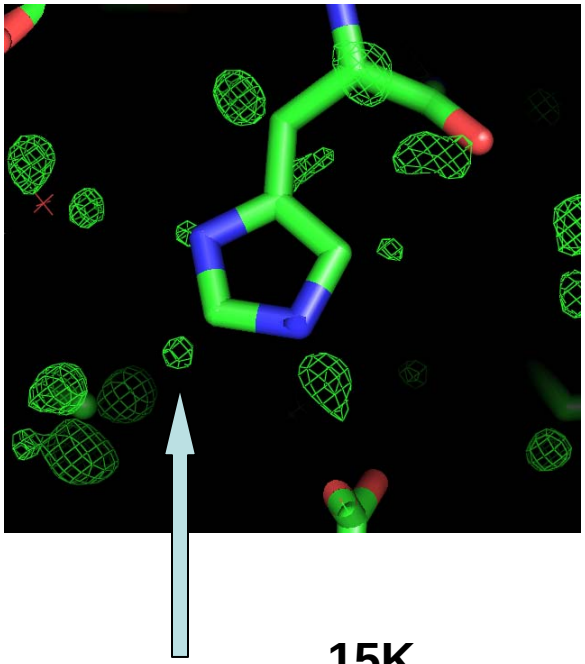
4.77 e/A3

Red Contours:

Fo-Fc

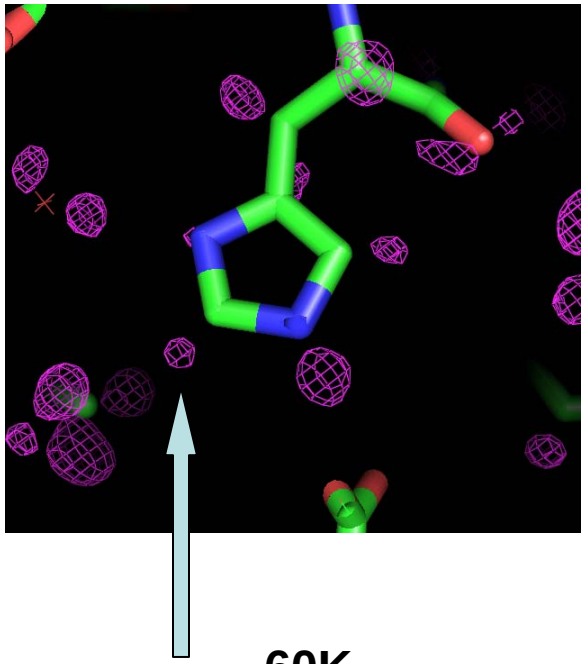
0.36 e/A3)

Hydrogen Atoms of Hys110 in the Active Site



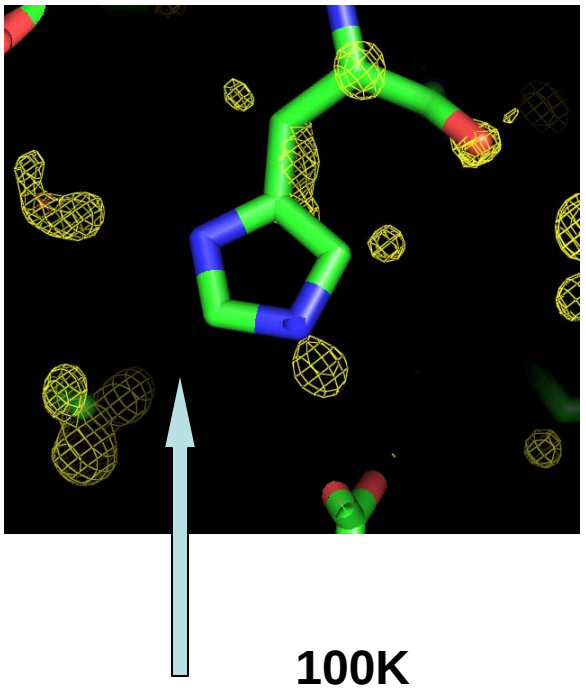
15K

Hydrogen atom



60K

Hydrogen atom

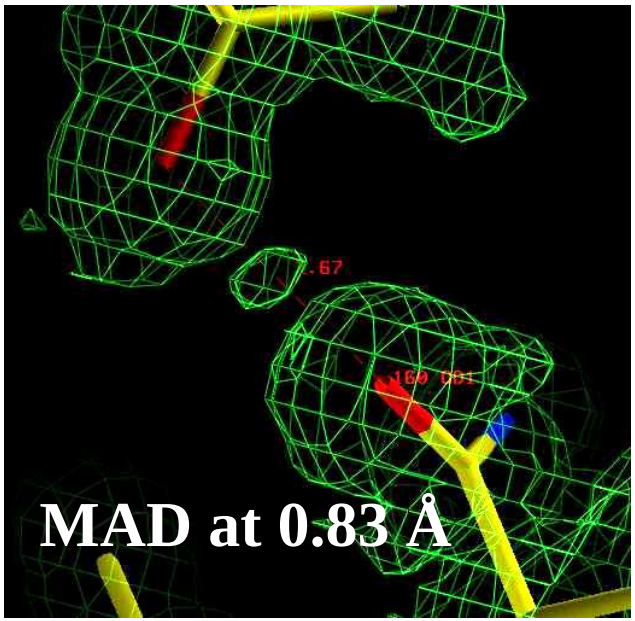


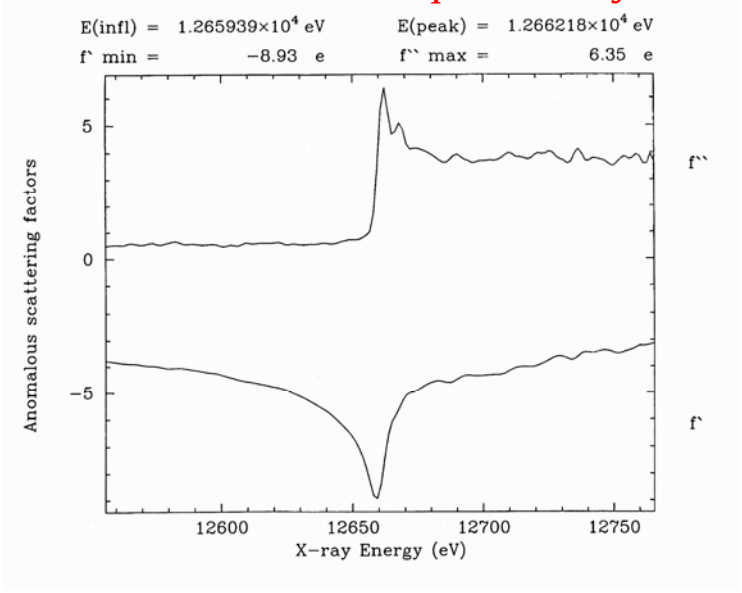
100K

Missing hydrogen atom

Multi- and Single-wavelength Anomalous Diffraction Using Synchrotron Sources Creates an Opportunity for Automation of Structure Determination

- All “heavy – N>50” and “light – 20<N<50” atoms show good anomalous signal associated with K, L and M edges
- “Heavy” atoms can be readily introduced into proteins (SeMet, Br, I, Xe, Ar, As, metal ions (Rb etc)) and DNA/RNA (Br)
- MAD/SAD does not require a native crystal
- Anomalous signal does not decay with resolution
- Use of anomalous signal simplifies approach to structure determination and improves isomorphism
- The anomalous signal is weak (1-6%)
- Optimal data collection requires a synchrotron facility

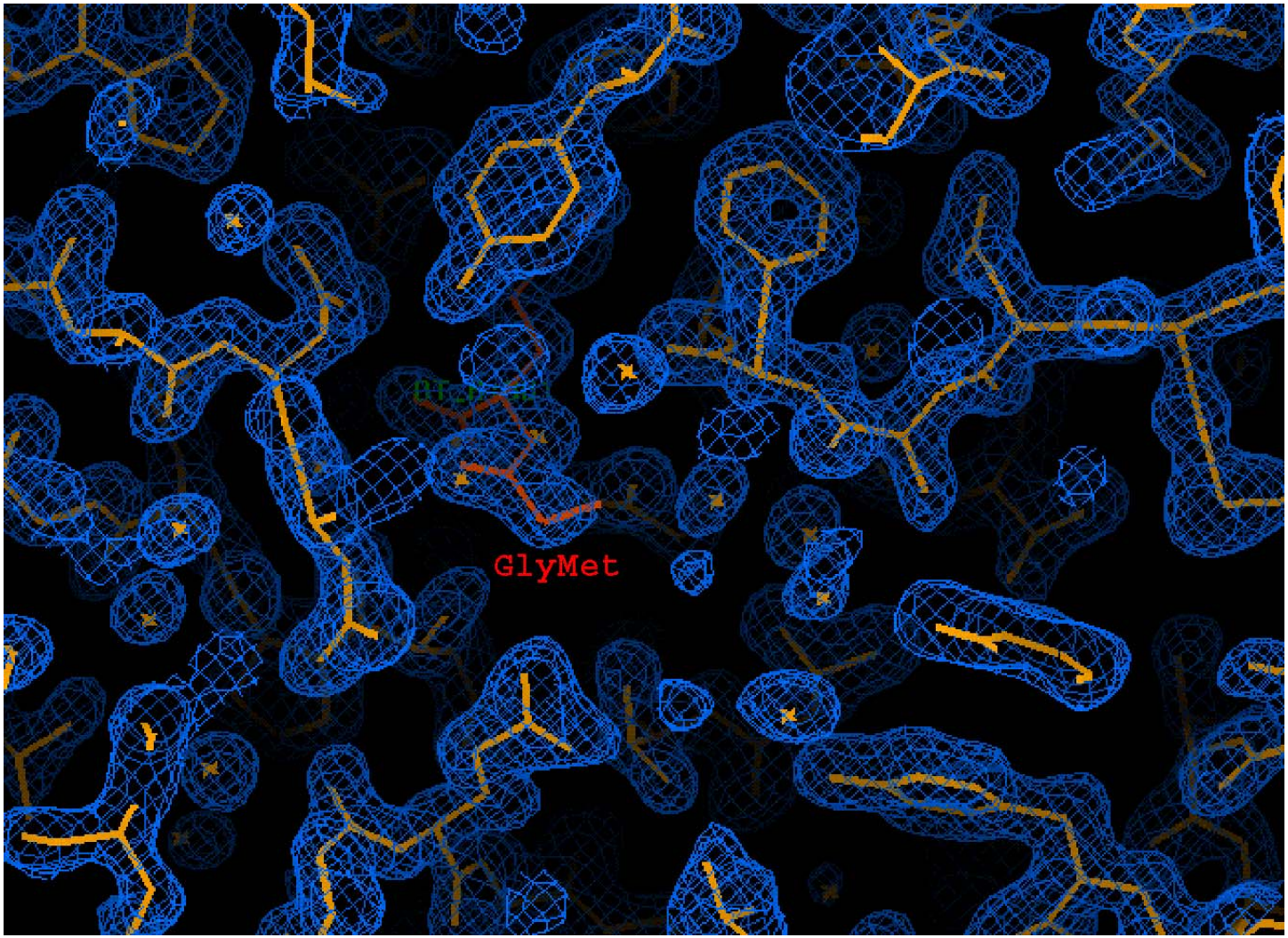




MAD/SAD for PROTEINS

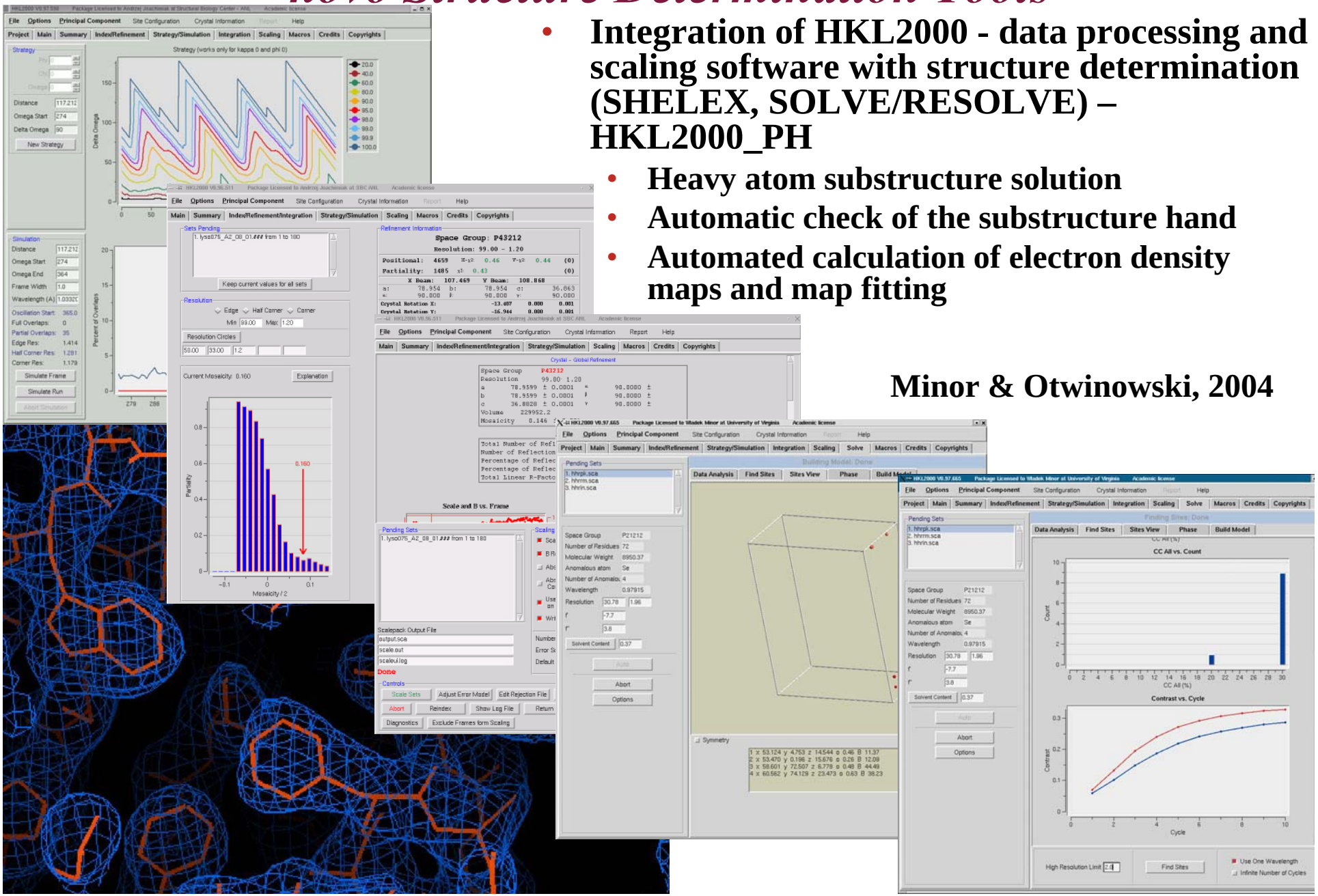
- *In vivo* protein labeling with SeMet
- Standard protocol for data collection and structure determination
- High-resolution and high quality allows auto-tracing

MAD/SAD Phasing Provides Higher Quality Electron Density Maps, Allows Automated Interpretation and Improves Structure Quality



APC009, 1.7 Å SAD Map 1σ, 1 Se/297 AA (32 kDa)

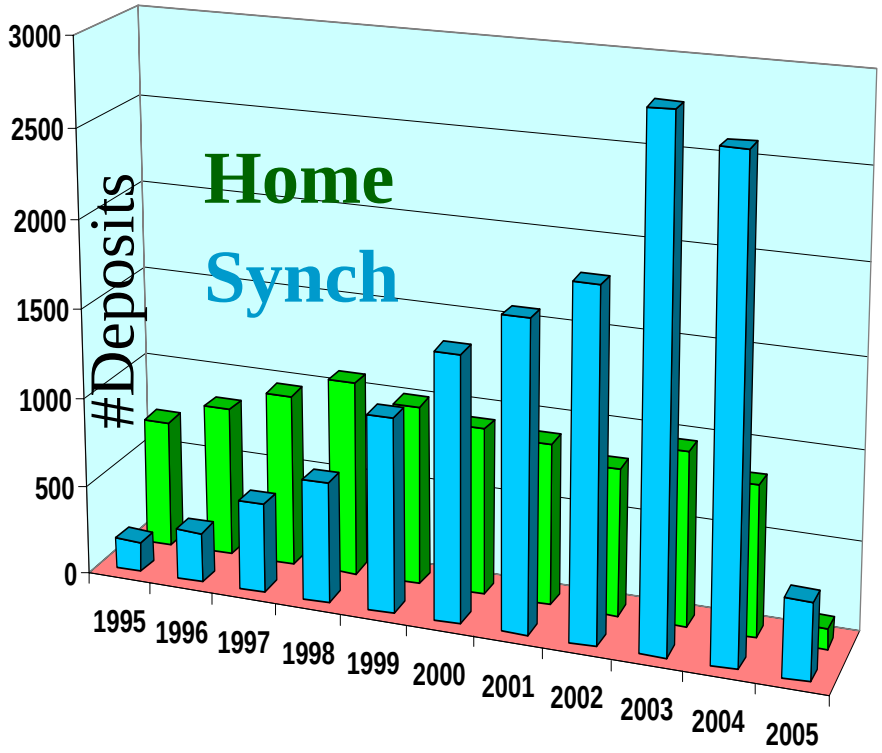
Integration of Data Collection Experiment with HTP de novo Structure Determination Tools

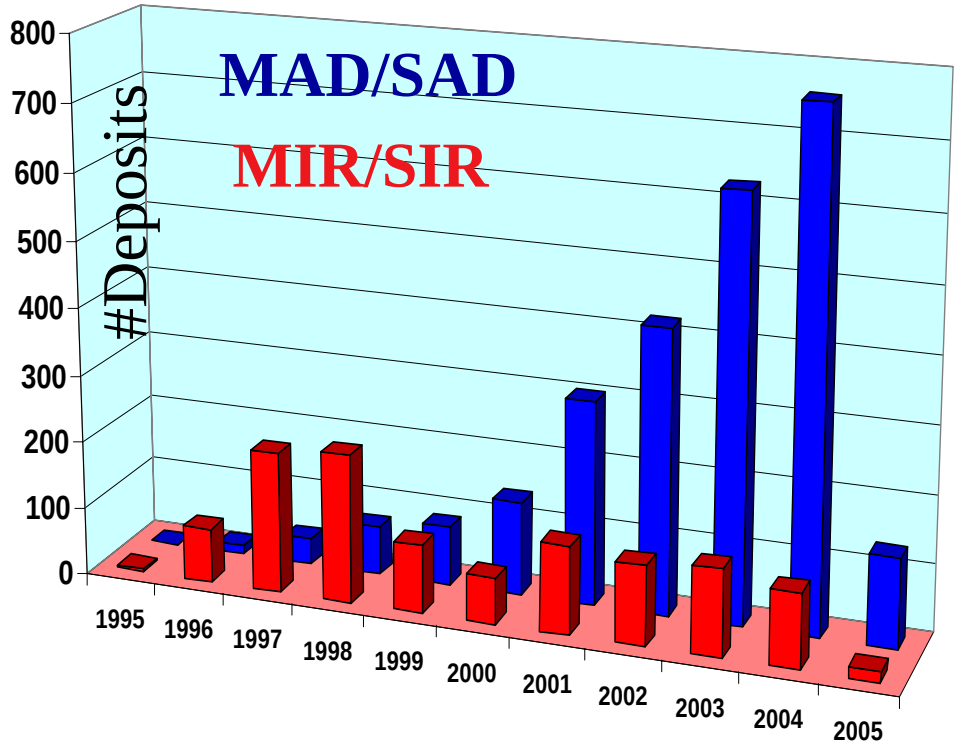


- **Integration of HKL2000 - data processing and scaling software with structure determination (SHELEX, SOLVE/RESOLVE) – HKL2000_PH**
  - **Heavy atom substructure solution**
  - **Automatic check of the substructure hand**
  - **Automated calculation of electron density maps and map fitting**

Minor & Otwinowski, 2004

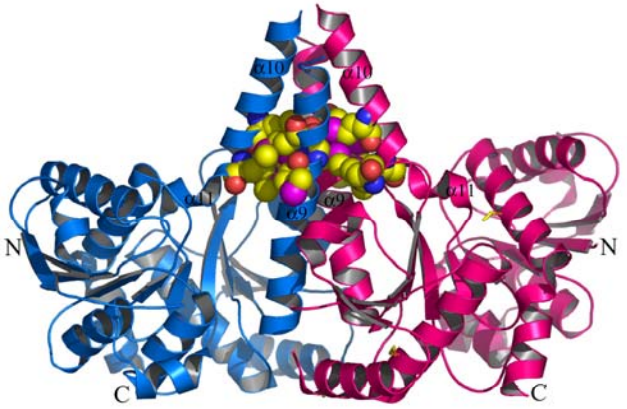
Effect of Synchrotron Sources on Protein Structure Determination

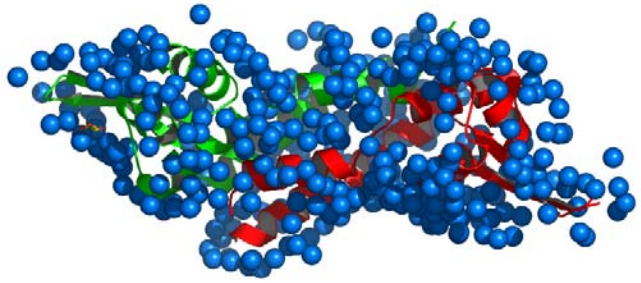


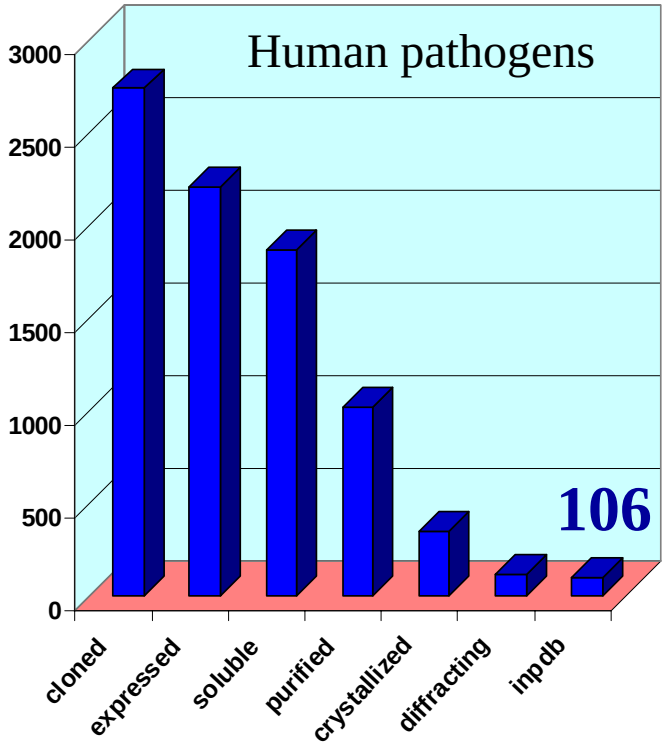


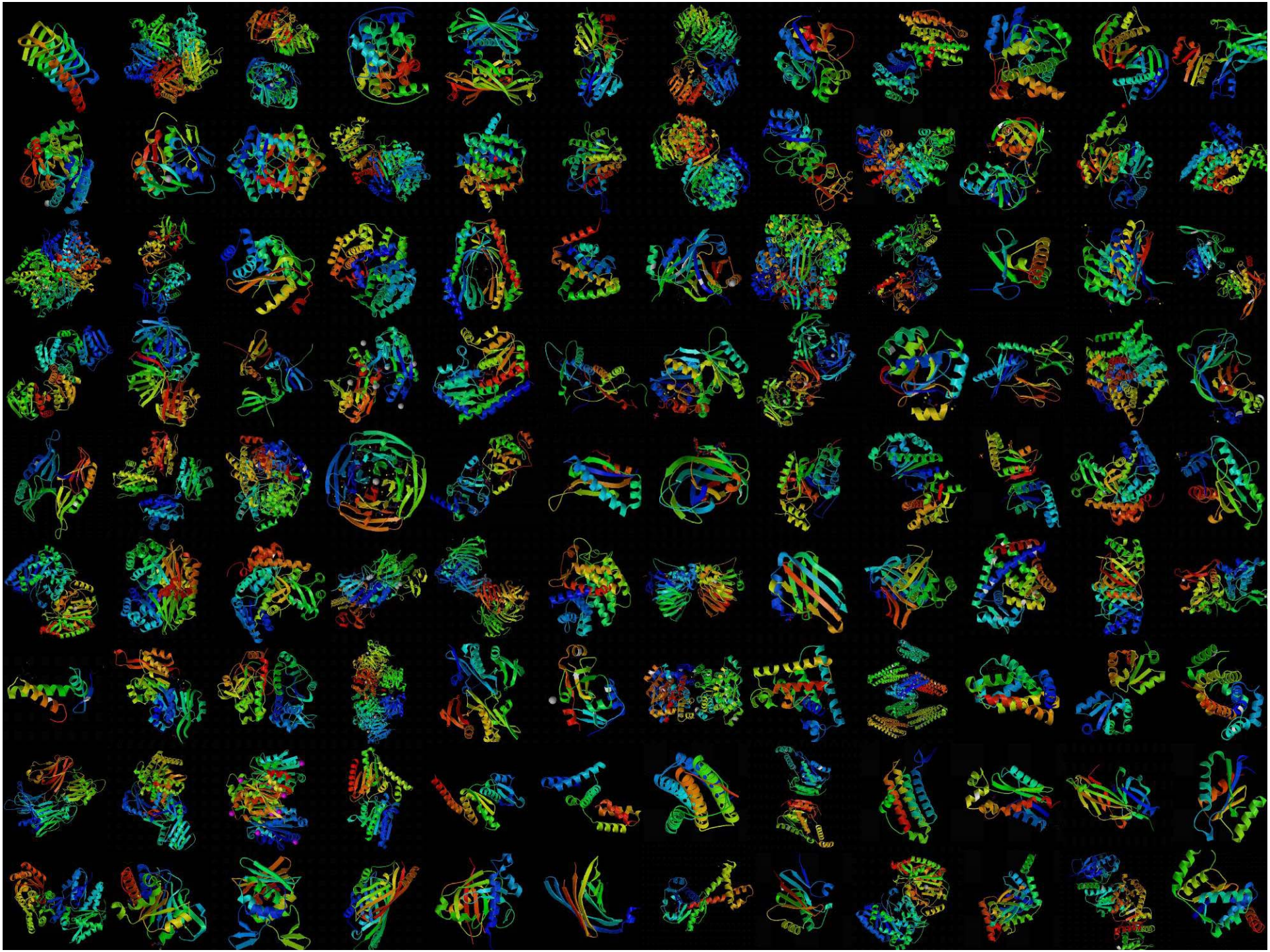
MCSG Progress, July 8, 2005

- **Targets (genes) 14,266**
- **Stopped projects 1,258**
- **Cloned genes 5,475**
- **Expressed proteins 4,052**
- **Soluble proteins 3,211**
- **Purified proteins 2,120**
- **Crystals 824**
- **Diffracting crystals 324**
- **New structures in PDB 273**
- **Total structures in PDB 281**
- **New folds 24**



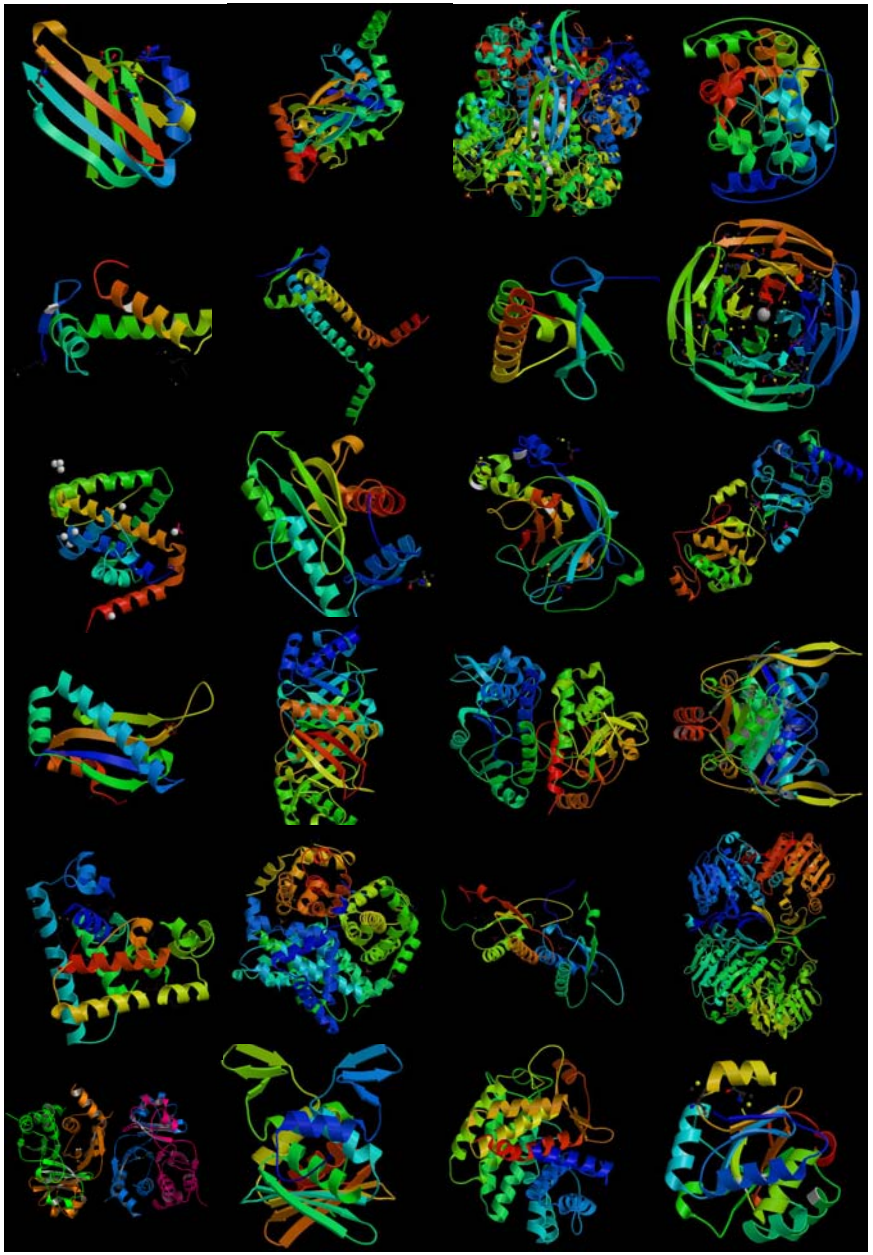






July 8, 2005 – 273 MCSG Structures in PDB

- **In 2004 MCSG has deposited 112 structures to PDB (103 in the past 10 months).**
- **Full lengths proteins:**
  - **Average gene size 319 AA (range 97 - 783AA).**
  - **MW~33 kDa (range 11 kDa – 90 kDa, 11kDa/AU - 330 kDa/AU).**
  - **On average proteins are dimers (range 1-16 subunits).**
  - **Average resolution 1.94 Å (range 1.1- 3.2 Å).**
  - **Average R=20.0%, Rfree=23.9%.**
  - **79% structures are unique**
- **All structures determined using anomalous signal and synchrotron radiation:**
  - **250 (91%) MAD/SAD with SeMet substituted proteins (range 1Se/297 – 108 Se/2720),**
    - **Six proteins were crystallized using mutagenesis (Derewenda)**
    - **Fifteen proteins were crystallized after reductive methylation**
  - **22 SAD/MIR with other HA (Hg, Pt, Zn, Cd, I).**
  - **1 molecular replacement.**



Security / Privacy Notice

MCSG

Midwest Center for Structural Genomics

PSI

Structure Galler

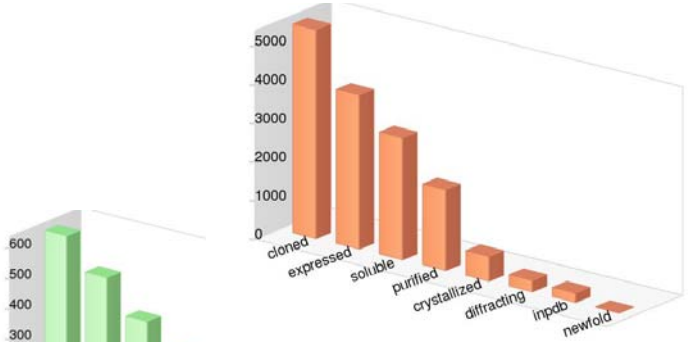
• XML Files • Target List • Progress • Statistics • Log in • Site Search: Go

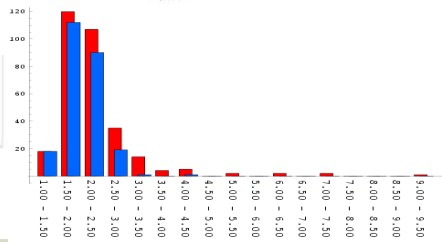
- Consortium
- Project
- Investigators
- Targets
- 3-D Structures
- Related Publications
- SG Sites
- SG Progress
- NIH
- MCSG Resources
- Job opportunities
- Collaborators
- Internals

Progress *as of Tuesday 12th of April 2005 01:04:27 PM*

Progress of the project, only started targets are displayed

MCSG ID	AccCode	Clone	Expression	Solubility	Purified	Crystals	Diffraction	Method	PDB ID	Status
APC006	NP_370689	+	+	+	+	+	+	MAD	1SQE	In PDB
APC007	NP_645835	+	+	+	+	+	+	MAD	1XBW	In PDB
APC008	BAB94840	+	+	+	+	+	+	MAD		Active
APC009	BAB41652	+	+	+	+	+	+	MAD	1P99	In PDB
APC010	BAB42231	+	+	+	+	+	+	MAD	1NG5	In PDB
APC011	P37650	+	+							Active
APC012	AAD36133	+	+	+	+	+	+	MAD	1KR4	In PDB
APC014	CAC11321	+	+	+	+	+	+	MAD	1KYT, 1L6R	In PDB
APC015	NP_813852	+	+	+	+					Stopped
APC019	Q9WZY7	+	+	+	+	+				Stopped-homolog solved by others
APC020	Q9WZZ9	+	+	+	+	+	+/-			Stopped
APC021	AAB90320	+	+	+	+					Active
APC022	RCJ00412	+	+	+	+	-				Stopped (failed trials)
APC023	NP_281856	+	+	+	+					Active

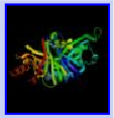




Target APC007

PDB 1XBW PDBSUM Profunc

Protein: isdG (ncbi, 107 aa) COG2329 related proteins



1XBW

Staphylococcus aureus
Deposited: 31 Aug 2004
Authors: Zhang, R., Wu, R., Joachimiak, G., Schneewind, O., Joachimiak, A.,

Target APC009

PDB 1P99 PDBSUM Profunc

Protein: SA0422 (ncbi, 280 aa) COG1464 related proteins

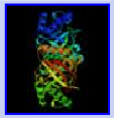
1P99

ABC transporter
Staphylococcus aureus
Deposited: 09 May 2003
Authors: Zhang, R., Zhou, M., Joachimiak, G., Schneewind, O., Joachimiak, A.,
references

Target APC010 - NEW FOLD

PDB 1NG5 PDBSUM Profunc

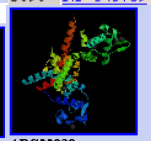
Protein: SA0982 (ncbi, 244 aa) COG4509 related proteins



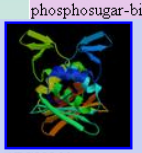
1NG5

Transpeptidase
Staphylococcus aureus
Deposited: 16 Dec 2002
Authors: Zhang, R., Joachimiak, G., Joachimiak, A.,
references

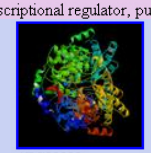
C079	G69065	ncbi	Putative protein	113	Methanobacterium thermoautotrophicum	In PDB
C080	P30747	ncbi,swprot	Molybdenum cofactor biosynthesis moaC protein	160	Escherichia coli	Stopped-homolog solved by others
C081	P08323	ncbi,swprot	Glycogen synthase	477	Escherichia coli	Active
C083	RBB00441	ncbi	Cell division protein ftsA	413	Borrelia burgdorferi	Stopped
C084	NP_344870	ncbi	cell division protein FtsL	105	Streptococcus pneumoniae	Active
C085	NP_345070	ncbi	conserved hypothetical protein	97	Streptococcus pneumoniae	In PDB
C086	NP_346101	ncbi	ylmH protein	261	Streptococcus pneumoniae	Active
C088	NP_346076	ncbi	Rrf2 family protein	145	Streptococcus pneumoniae	Active
C089	RCJ00595	wit	Transcription antitermination protein NUSG	177	Campylobacter jejuni	Active
C090	RCJ00223	wit	Hypothetical transcriptional regulator	221	Campylobacter jejuni	Active
C091	RCJ00475	wit	Transcriptional regulatory protein HYPF	729	Campylobacter jejuni	Active
C092	RNG00001	ncbi	Transcription termination factor RHO	213	Neisseria gonorrhoeae	Stopped-homolog solved by MCSG
C093	NP_345223	ncbi	conserved hypothetical protein	230	Streptococcus pneumoniae	Active
C094	NP_346330	ncbi	msm operon regulatory protein	286	Streptococcus pneumoniae	Active
C095	NP_344919	ncbi	transcriptional regulator, putative	651	Streptococcus pneumoniae	Active
C096	NP_345789	ncbi	phosphosugar-binding transcriptional regulator, putative	269	Streptococcus pneumoniae	Active



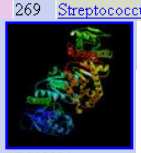
APC22838
1YLF **ident: 41.1%**
annotation



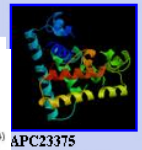
APC22846
1R7L New Fold
annotation



APC22880
1XR4 **ident: <20%**
annotation



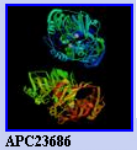
APC22886
1R0K **ident: 92.5%**
annotation



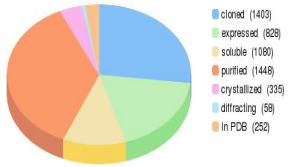
APC23375
1S4K New Fold
annotation

APC23398
1S9U **ident: 21.6%**
annotation

APC23655
1Y7R **ident: <20%**
annotation



APC23686
1XV2 **ident: <20%**
annotation



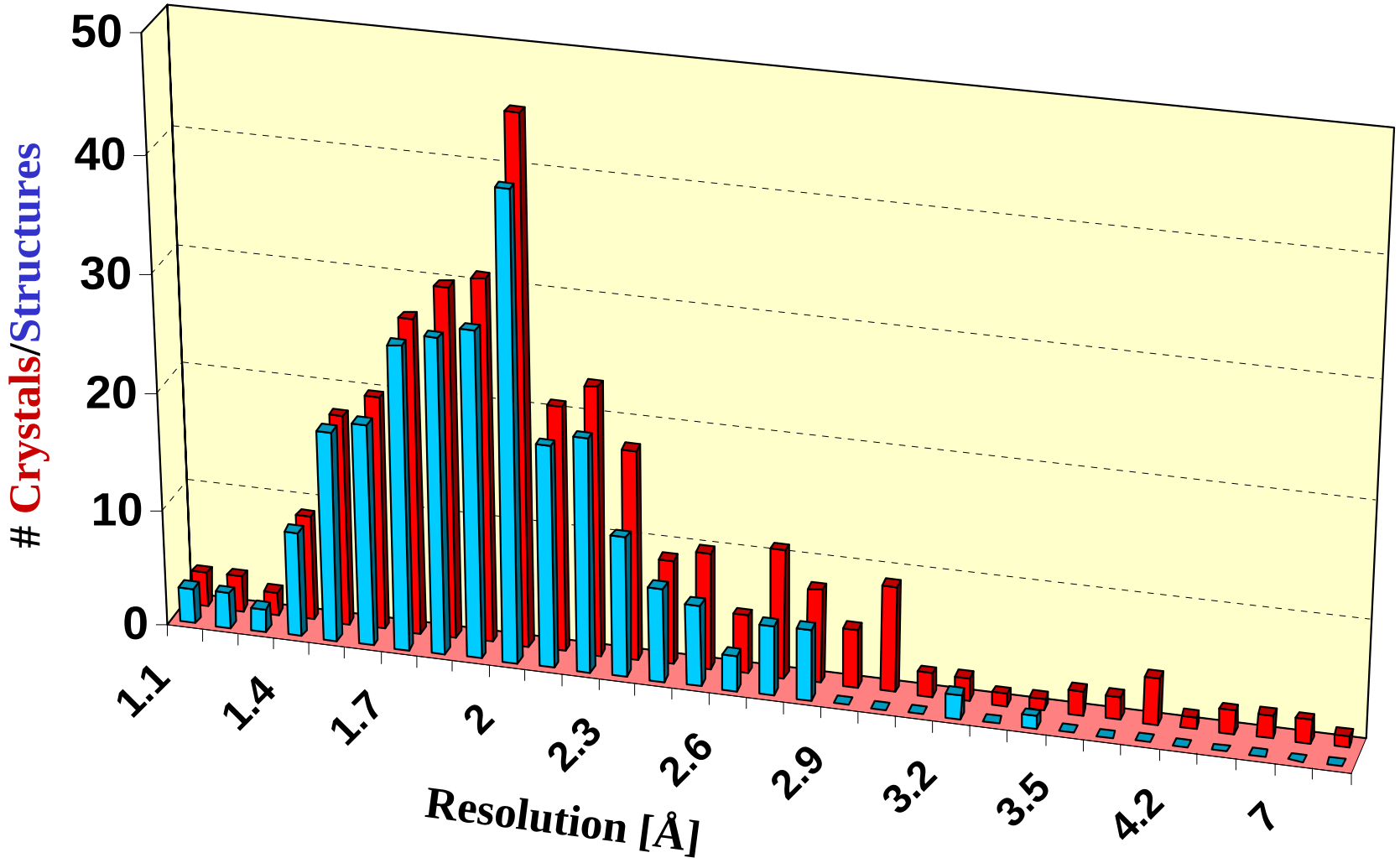
5. **APC006** Wu R, Skaar EP, Zhang R, Joachimiak G, Gornicki P, Schneewind O, Joachimiak (2004)
Staphylococcus aureus IsdG and IsdI, heme degrading enzymes with structural
J Biol Chem, in press. [PubMed] [PDB]

6. **APC012** Savchenko A, Skarina T, Evdokimova E, Watson JD, Laskowski R, Arrowsmith CH, Edwards AM, Joachimiak A, Zhang RG (2004)
X-ray crystal structure of CutA from Thermotoga maritima at 1.4 A resolution.
Proteins, 54, 162-5. [PubMed] [PDB]

7. **APC014** Kim Y, Yakunin AF, Kuznetsova E, Xu X, Pennycooke M, Gu J, Cheung F, Proudfoot M, Arrowsmith CH, Joachimiak A, Edwards AM, Christendat D (2004)
Structure- and function-based characterization of a new phosphoglycolate phosphatase from Thermoplasma acidophilum.
J Biol Chem, 279, 517-26. [PubMed] [PDB] [PDB]

8. **APC1138** Rajan SS, Yang X, Collart F, Yip VL, Withers SG, Varrot A, Thompson J, Davies GJ, Anderson WF (2004)
Novel Catalytic Mechanism of Glycoside Hydrolysis Based on the Structure of an NAD(+)/Mn(2+)-Dependent Phospho-alpha-Glucosidase from Bacillus subtilis.
Structure (Camb), 12, 1619-29. [PubMed] [PDB]

~13% of MCSG Structures Have Been Determined at 1.5 Å or Higher Resolution

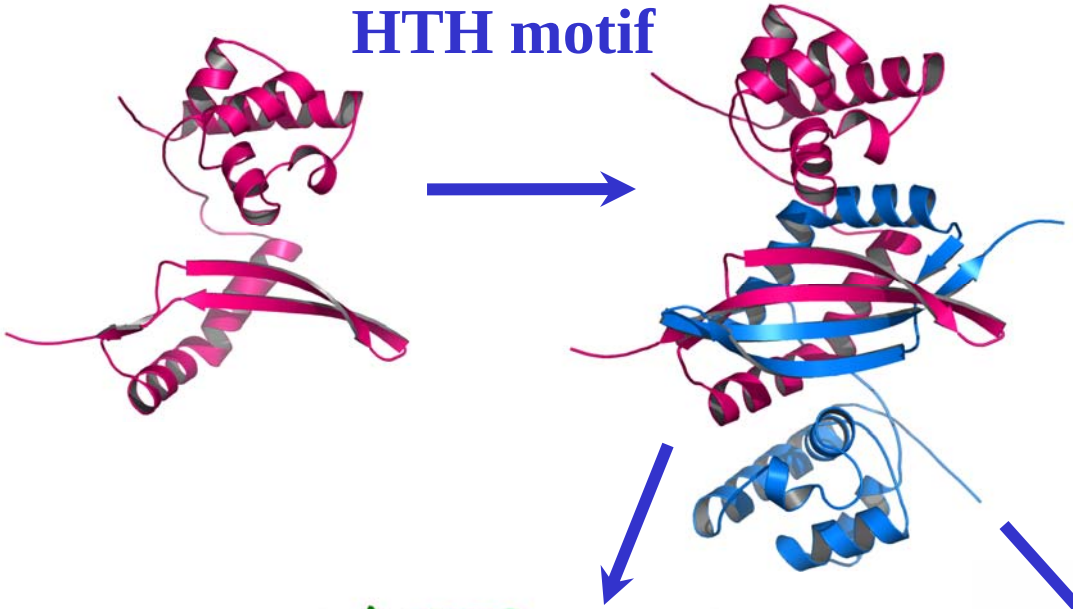


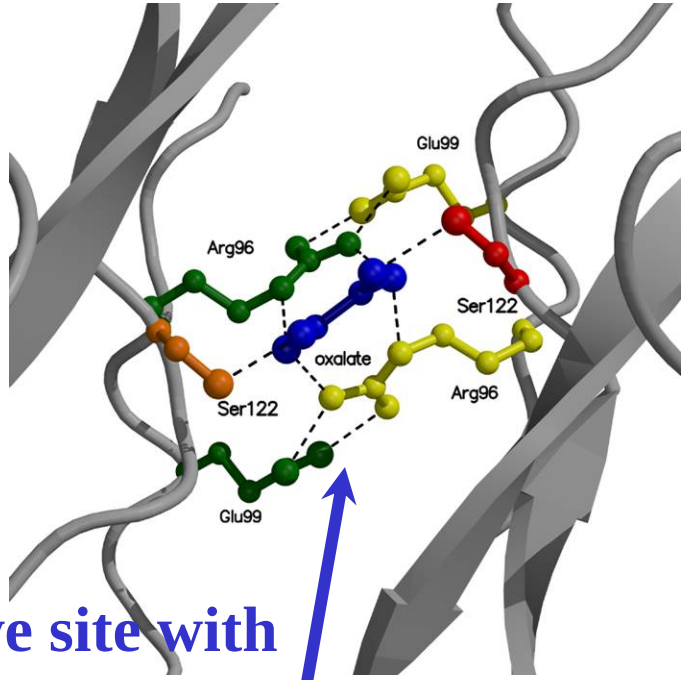
Structural Genomics can Contribute at Least Three Tools

- **A comprehensive dictionary of high-resolution protein structures determined experimentally by x-ray crystallography and NMR**
- **A comprehensive library of recombinant protein expression clones representing protein structures and functions**
- **Methods for automated, HTP implementation of the currently powerful but tedious and labor-intensive protocols of molecular biology**
- **Some functional information derived from structure**

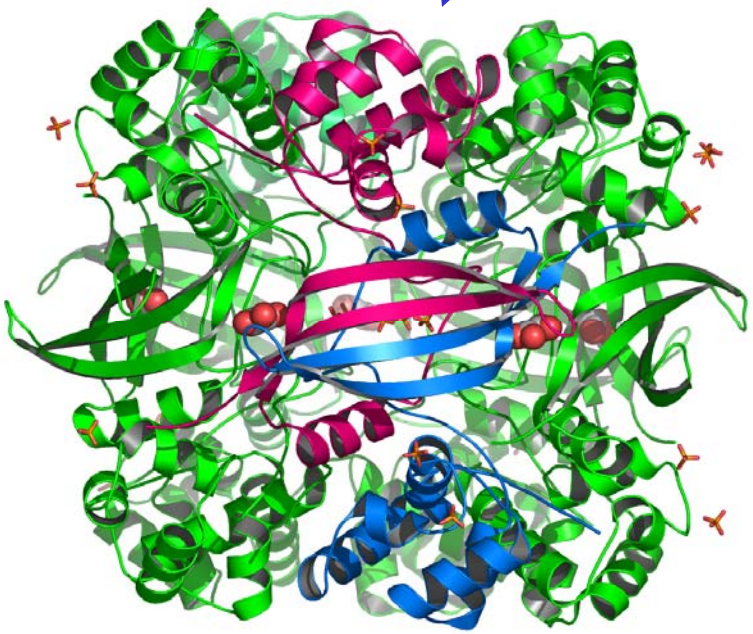
Cyanase – an Enzyme Created by Oligomerization of DNA-Binding Protein

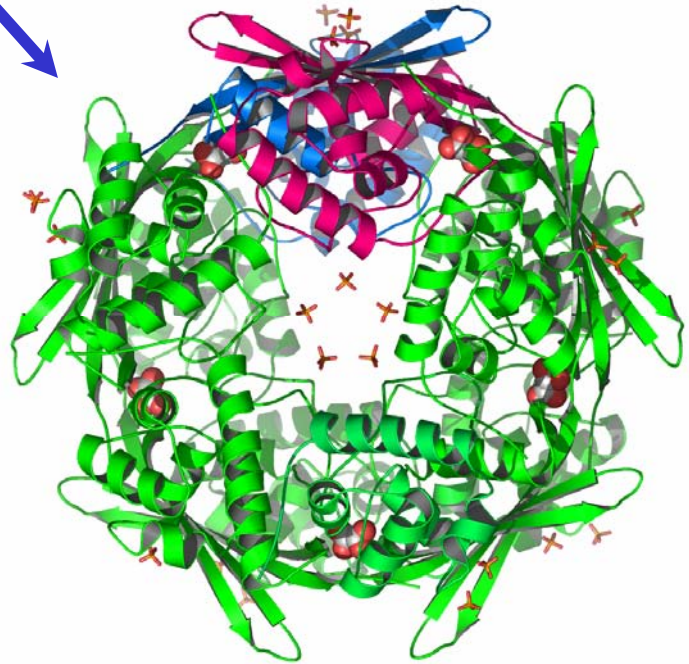
HTH motif





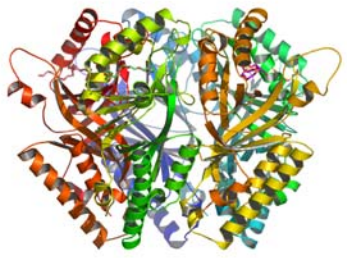
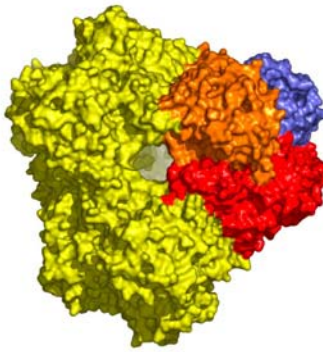
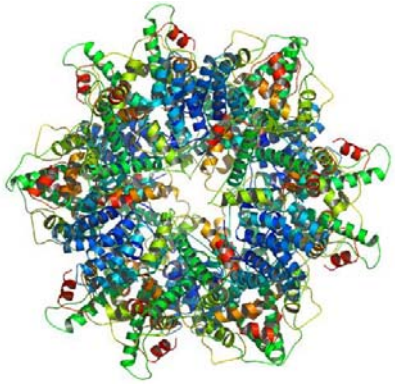
Active site with bound oxalate

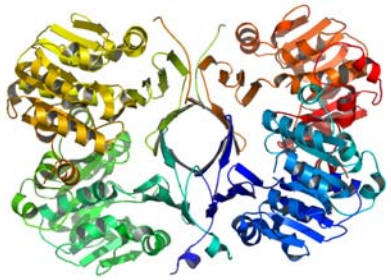
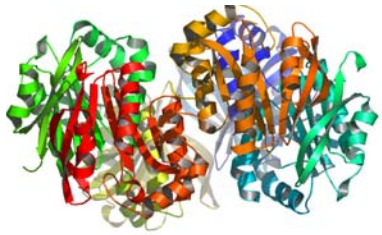




Difficult Structures



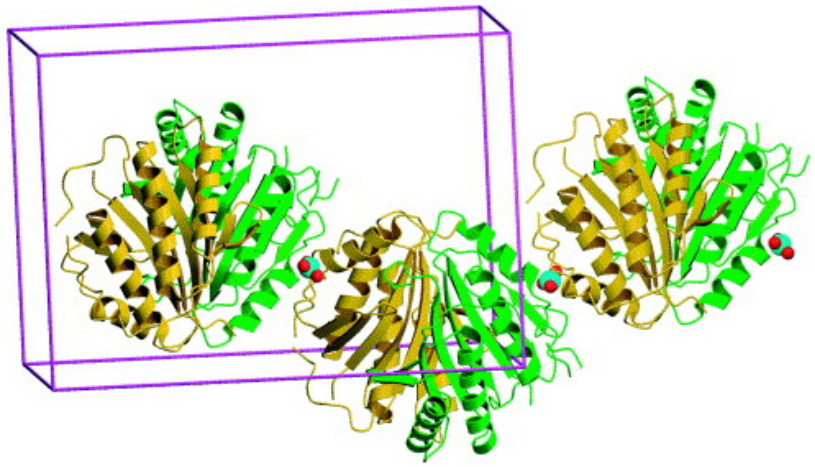





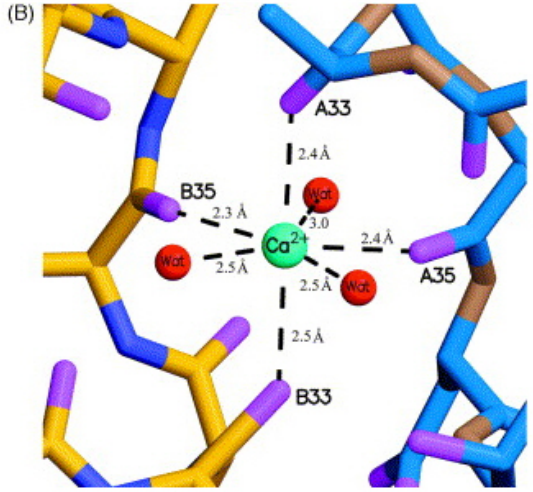



APC #	AA/Subunit	Sub/AU	AA/AU	# HA/AU	R (%)	Rfree (%)	Res (Å)
042	170	16	2,720	Hg-32	19.9	24.7	2.0
27836	345	6	2,070	Se-72	20.3	23.5	2.15
35865	249	8	1,992	Se-36	14.5	17.5	1.7
5057	147	12	1,764	Se-12	20.7	26.9	2.5
127	160	10	1,600	Se-40	14.6	18.1	1.7
24638	186	8	1488	Se-32	In prog.	In prog.	2.3
22852	164	9	1,476	Se-18	22.5	27.0	2.5
172	358	4	1,432	Se-36	23.3	26.5	2.0
4470	343	4	1,372	Se-36	23.9	29.8	2.6
26686	337	4	1,348	Se-64	In prog.	In prog.	2.35
35880	248	5	1,240	Se-40	15.8	19.4	1.75
5046	310	4	1,240	Se-28	14.3	17.8	1.6
047	296	4	1,184	Se-40	19.9	21.3	1.5
24328	292	4	1,168	Se-12	19.2	24.0	1.8
1068	184	6	1,104	Se-18	25.6	27.8	2.7
5029	134	8	1,072	Se-56	20.0	23.0	1.7
23620	292	4	1,168	Se-24	23.0	27.0	2.0
22880	509	2	1018	Se-24	19.3	23.5	2.4
50001	783	1	783	Se-17	19.7	23.6	2.0

Surface Conformational Entropy Reduction

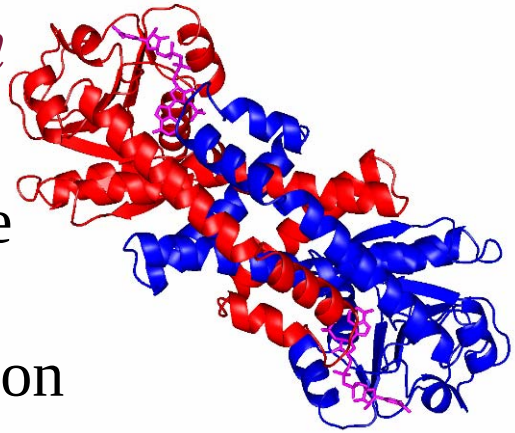
- A surface conformational entropy reduction was applied to a set of 20 *B. subtilis* proteins that failed in the initial crystallization screens.
- Two double mutants of YkoF - a thiamin-binding protein - (K112A/E114A and K33A/K34A) were designed. The K33A/K34A mutant crystallized readily and showed a strong requirement for divalent ions.
- The crystal structure was determined to 1.65 Å.
- The mutagenesis created an intermolecular Ca^{2+} binding site, essential for the formation of the crystal lattice.
- From a set of 20 *B. subtilis* targets we were able to crystallize 11 proteins, 7 as native and 6 mutated (2 proteins crystallize both as native and mutated).





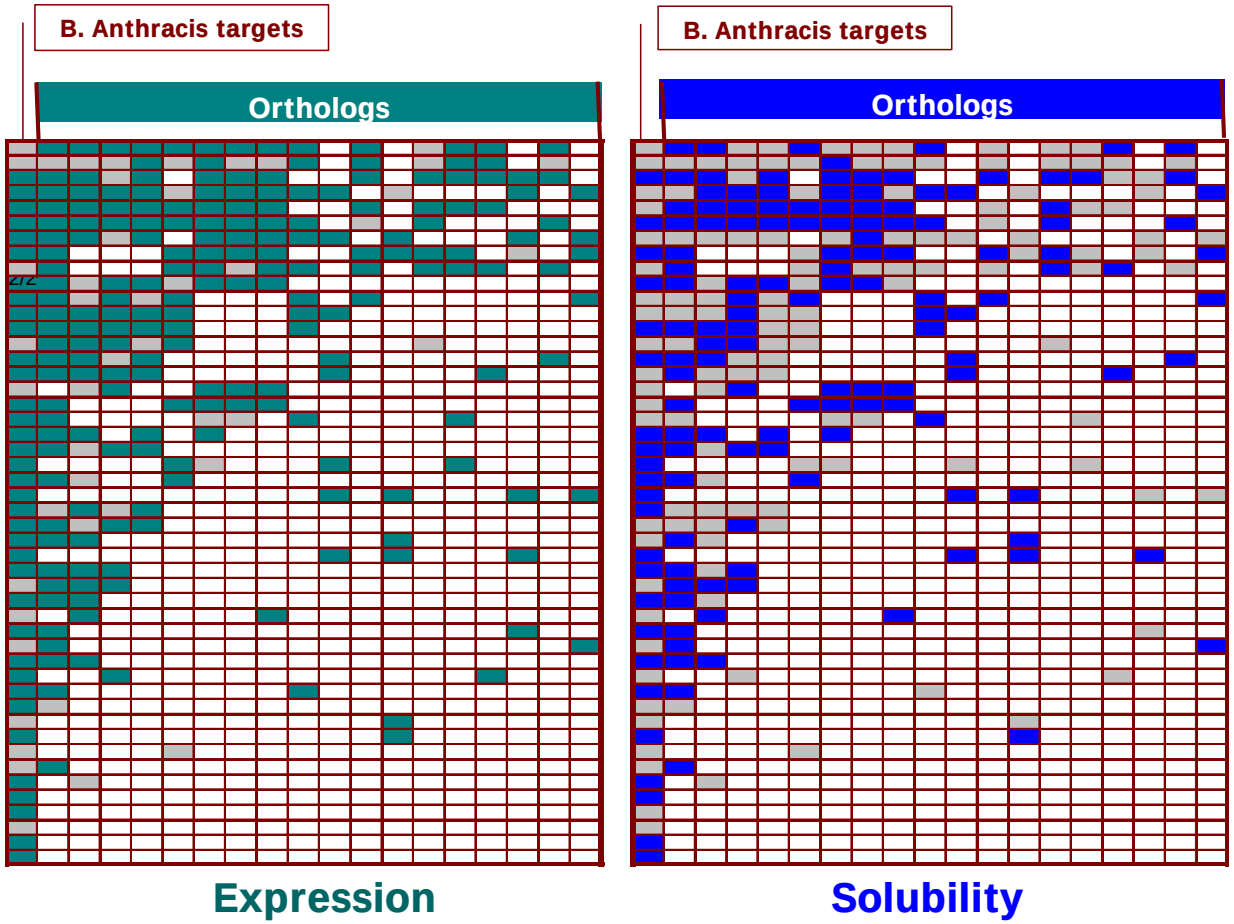
Derewenda et al.

Orthologue Scanning Approach



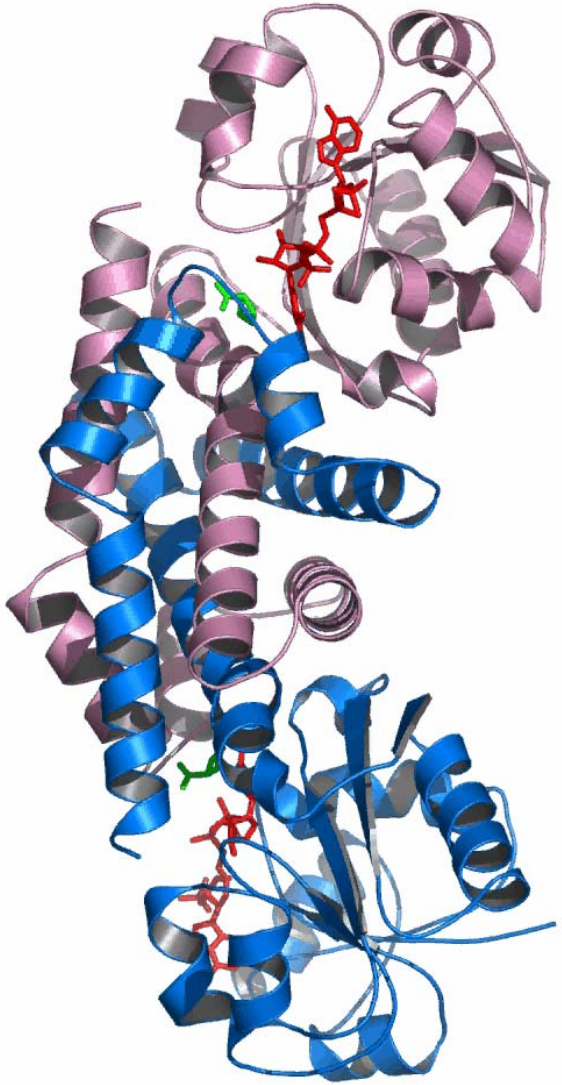
- 48 presumed essential genes in *B. anthracis* were used to identify orthologues in 18 genomes.
- 233 genes were cloned and screened for expression and solubility.

Organism	Strain	GC Con
Enterococcus faecalis	V583	37.53%
Streptococcus pyogenes	SF370	38.51%
Pyrococcus horikoshii	JCM 9974	41.88%
Bacillus subtilis	168	43.52%
Shewanella oneidensis	MR-1	45.96%
Archaeoglobus fulgidus	DSM 4304	48.58%
Escherichia coli	W1485	50.79%
Aeropyrum pernix	JCM 9820	56.31%
Haemophilus influenzae	Rd [KW20]	38.15%
Helicobacter pylori	J99	39.19%
Streptococcus pneumoniae	TIGR4	39.60%
Pyrococcus furiosus	DSM 3638 [Vc1]	40.77%
Vibrio cholerae	N16961	47.69%
Porphyromonas gingivalis	W83	48.29%
Shigella flexneri	2457T	50.80%
Neisseria meningitidis	MC 58	51.40%
Salmonella choleraesuis	LT2	52.22%
Thermoplasma acidophilum	DSM 1728	45.99%



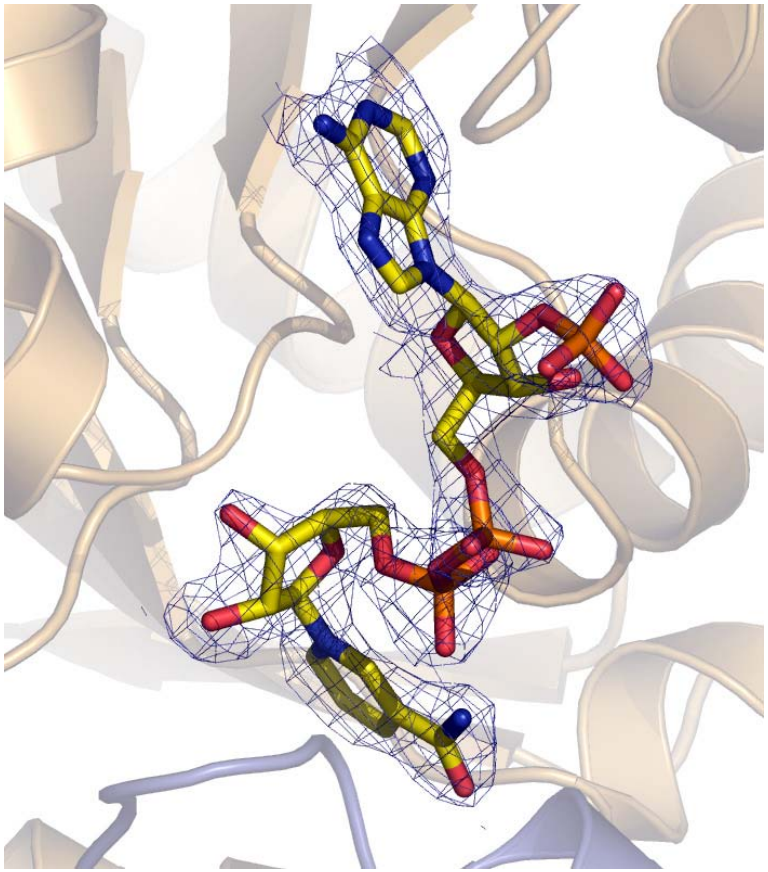
Crystal Structures of Δ1-Pyrroline-5-Carboxylate Reductase from Neisseria meningitides and Streptococcus pyogenes

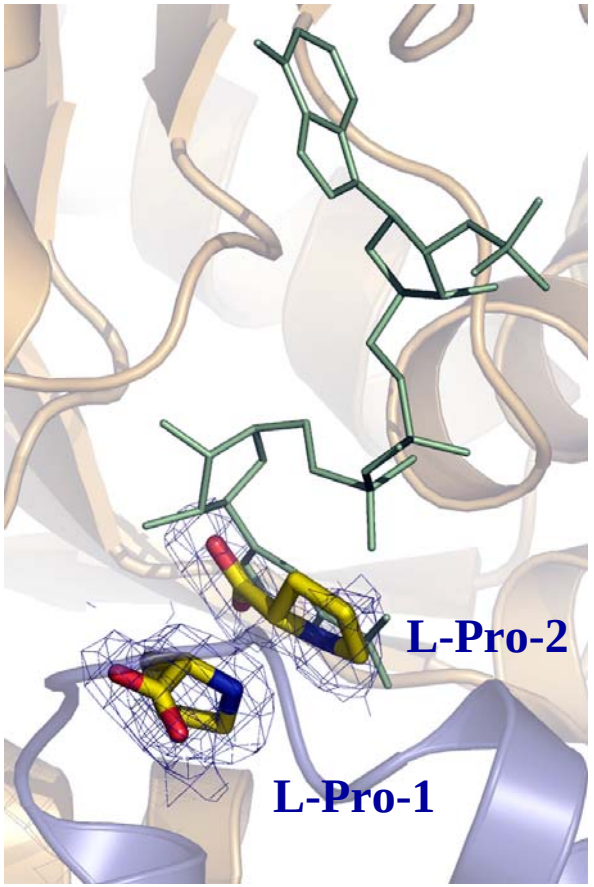
- **L-proline plays an important role in proteins**
- **The last step of proline biosynthesis, the conversion of Δ1-pyrroline-5-carboxylate (P5C) to L-proline is catalyzed by Δ1-pyrroline-5-carboxylate reductase (P5CR) using NADPH as cofactor**
- **P5CR is a member of a very large family (>400 family members)**
- **To increase the chance of obtaining a crystal structure from a member of this large and important family of enzymes, orthologues from 14 organisms have been cloned.**
- **9 P5CR enzymes expressed, 3 crystallized and 2 produced x-ray quality crystals**
- **The catalytic unit of P5CR is a dimer composed of "three" functional domains**
- **The N-terminal domain of P5CR shows an alpha-beta-alpha sandwich – a Rossman fold**
- **The C-terminal dimerization domain is rich in α-helical structure and shows domain swapping**



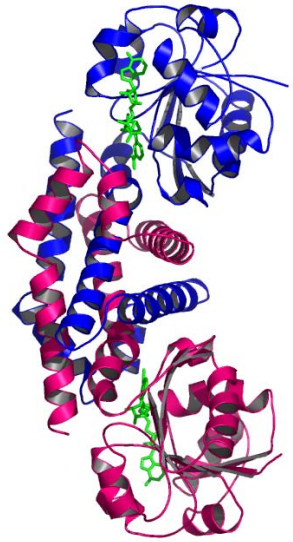
Binding $NADP^+$ and L-Proline

- Comparison of the P5CR native structure with structures P5CR complexed with L-proline and $NADP^+$ provides unique information about key functional features, the active site and the catalytic mechanism
  - Each ligand is bound to a distinct domain
  - Hydrite is transferred from pro-S face of nicotinamide ring and is unidirectional
  - Crystal structure shows how product L-Pro inhibits reaction

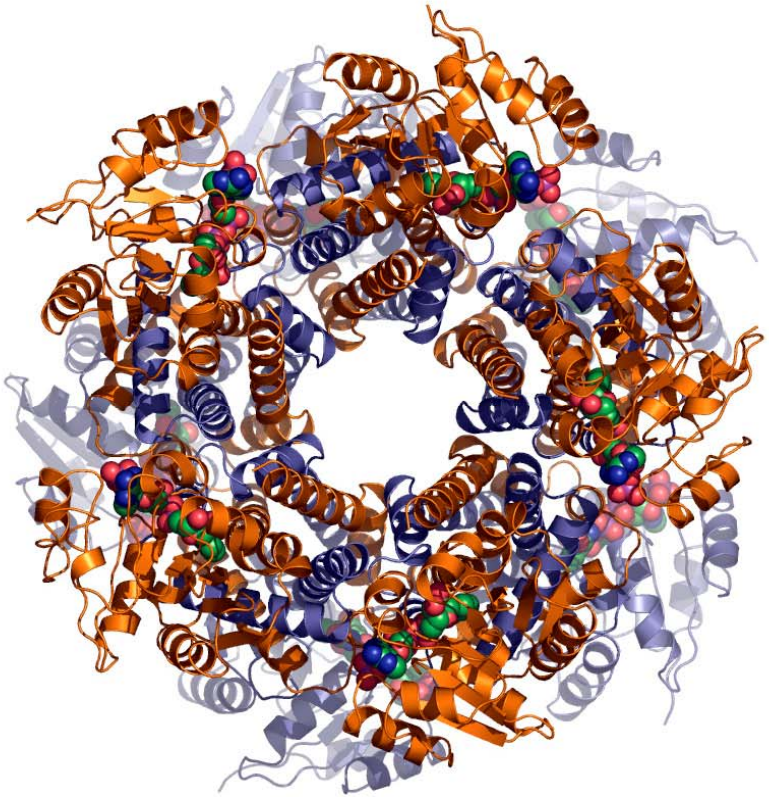


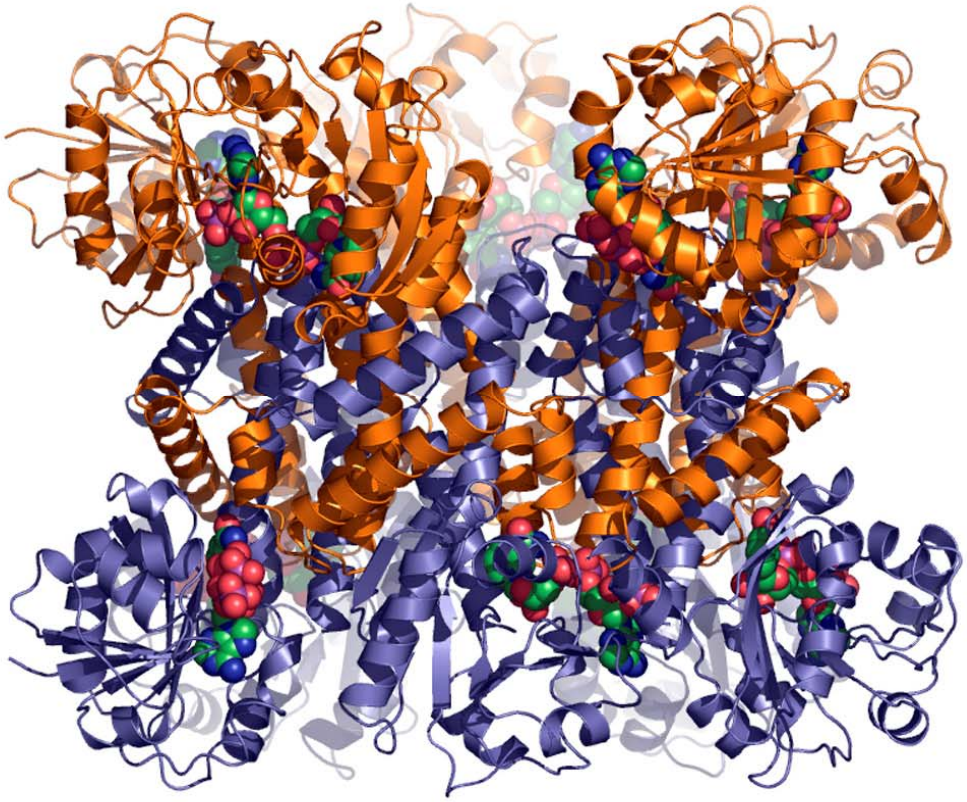


P5CR Basic Catalytic Unit is a Dimer, but it can Assemble into Larger Structures that Seems Species Specific



- P5CR from *N. meningitides* is a dimer
- P5CR from *S. pyogenes* is a decamer (pentamer of dimers)

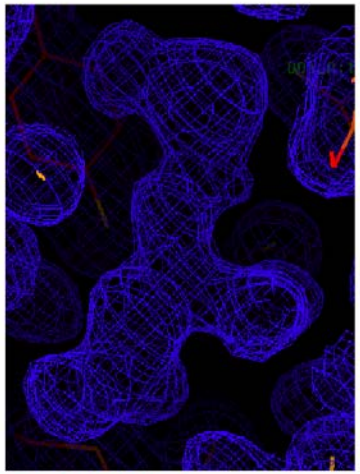


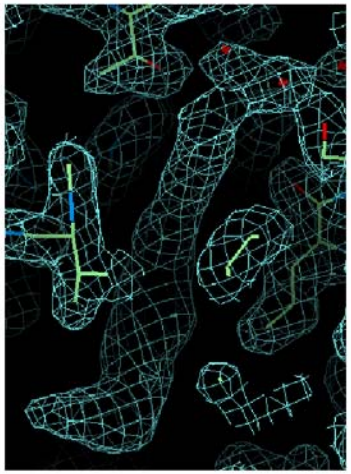


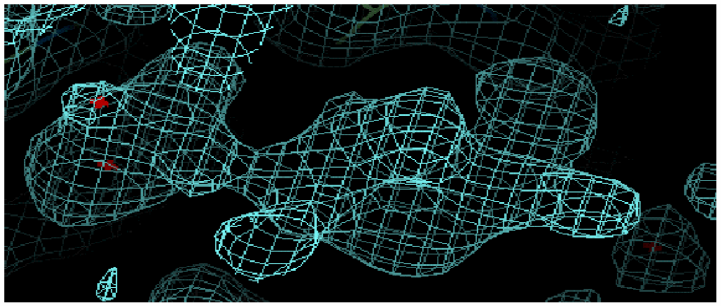
Structural Genomics Results

- **New structures confirmed the hypothesis that the structure-based classification of proteins contains far fewer protein families than sequence-based classifications:**
  - **Protein structure is better conserved than amino acid sequence, and**
  - **Can reveal distant evolutionary relationships that are undetectable by sequence comparisons**
- **Structures of singletons and proteins from very small families showed familiar folds, contradicting the hypothesis that these families may represent a rich reservoir of new folds**
  - **Sequence-based methods have major limitations for identifying proteins with potentially new folds**

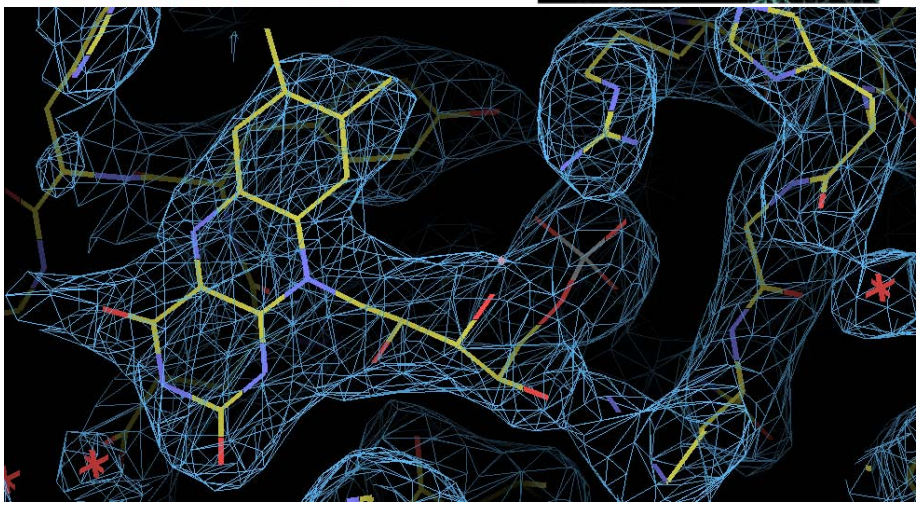
Ligands Found in some of MCSG Structures





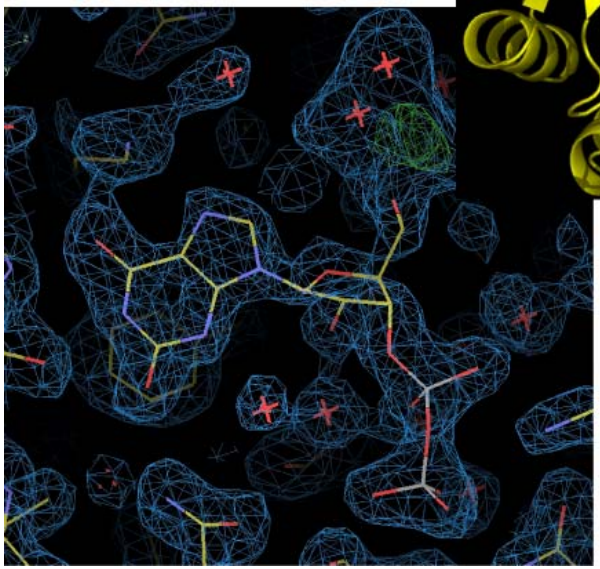


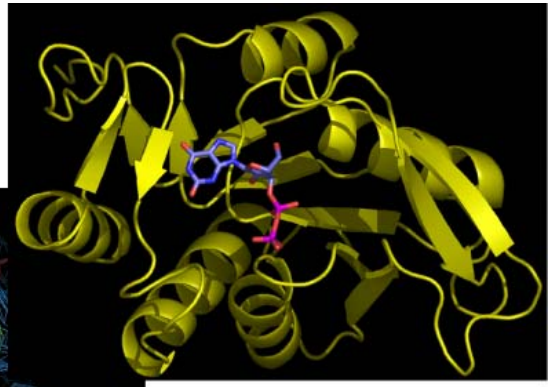
densities of the ligand in APC24838



APC1167

APC1452





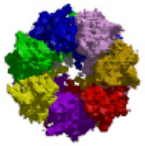
XANTHINE
PHOSPHORIBOSYLTRANSFERASE

from *Bacillus subtilis*

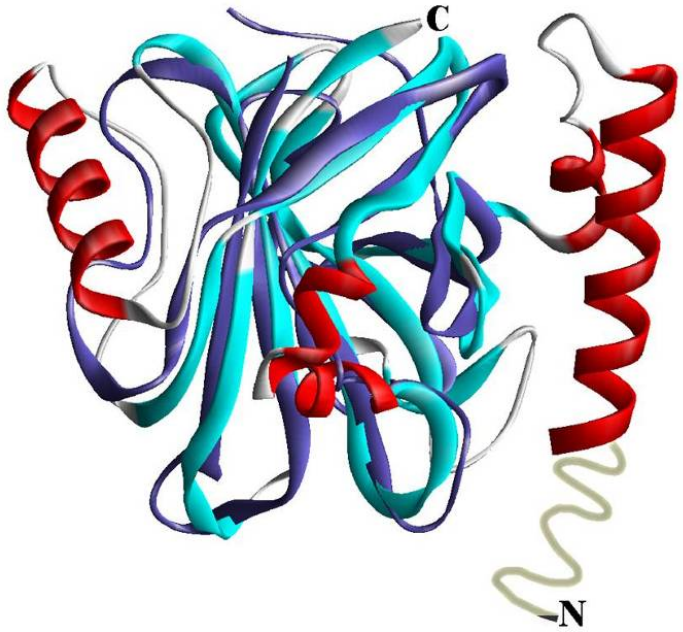
refinement in progress

R.Wu
M. Cuff
MCSG-ANL 2004

Functional Analysis



ProFunc *Analysis of a protein's 3D structure to help identify its likely biochemical function*



The aim of the ProFunc server is to help identify the likely biochemical function of a protein from its three-dimensional structure. It uses a series of methods, including fold matching, residue conservation, surface cleft analysis, and functional 3D templates, to identify both the protein's likely active site and possible homologues in the PDB.

Some of the methods take minutes to run; others take hours. You will be notified by e-mail when the entire process is complete, but can check on preliminary results as they become available.

From this page you can submit your own structure, analyse an existing PDB entry, or retrieve the results of a previously submitted run.

Choose option A, B or C:

A

Upload PDB-format file: Browse... UPLOAD

or

B

Get existing PDB file PDB code:

or

C

Go to previous analysis Id no.:

ProFunc

SSM results

Structural homologues

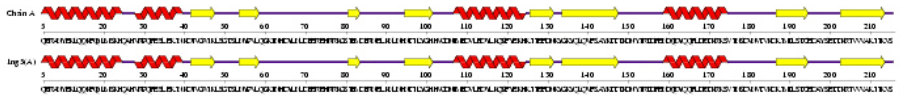
The table below shows the 1 structural match returned by the SSM program.

To view any of the matched structures superposed on the query structure in RasMol, use the checkboxes to select the structrues to view and press the View button

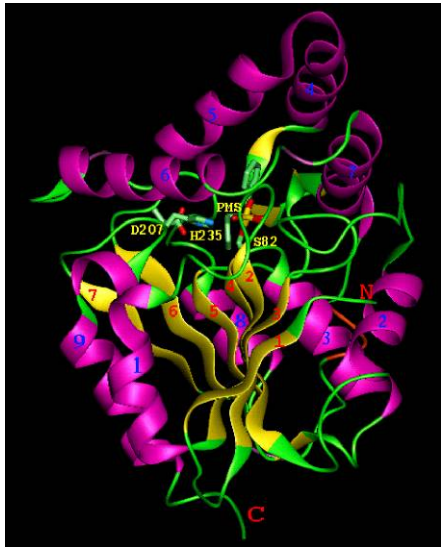
Match	Select	Z-score	No. SSE	RMSD	Seq. id	PDB entry	Name
1.	☑	14.9	13	0.00Å	100.0%	1ng5(A)	2.0 a crystal structure of staphylococcus aureus sortase b

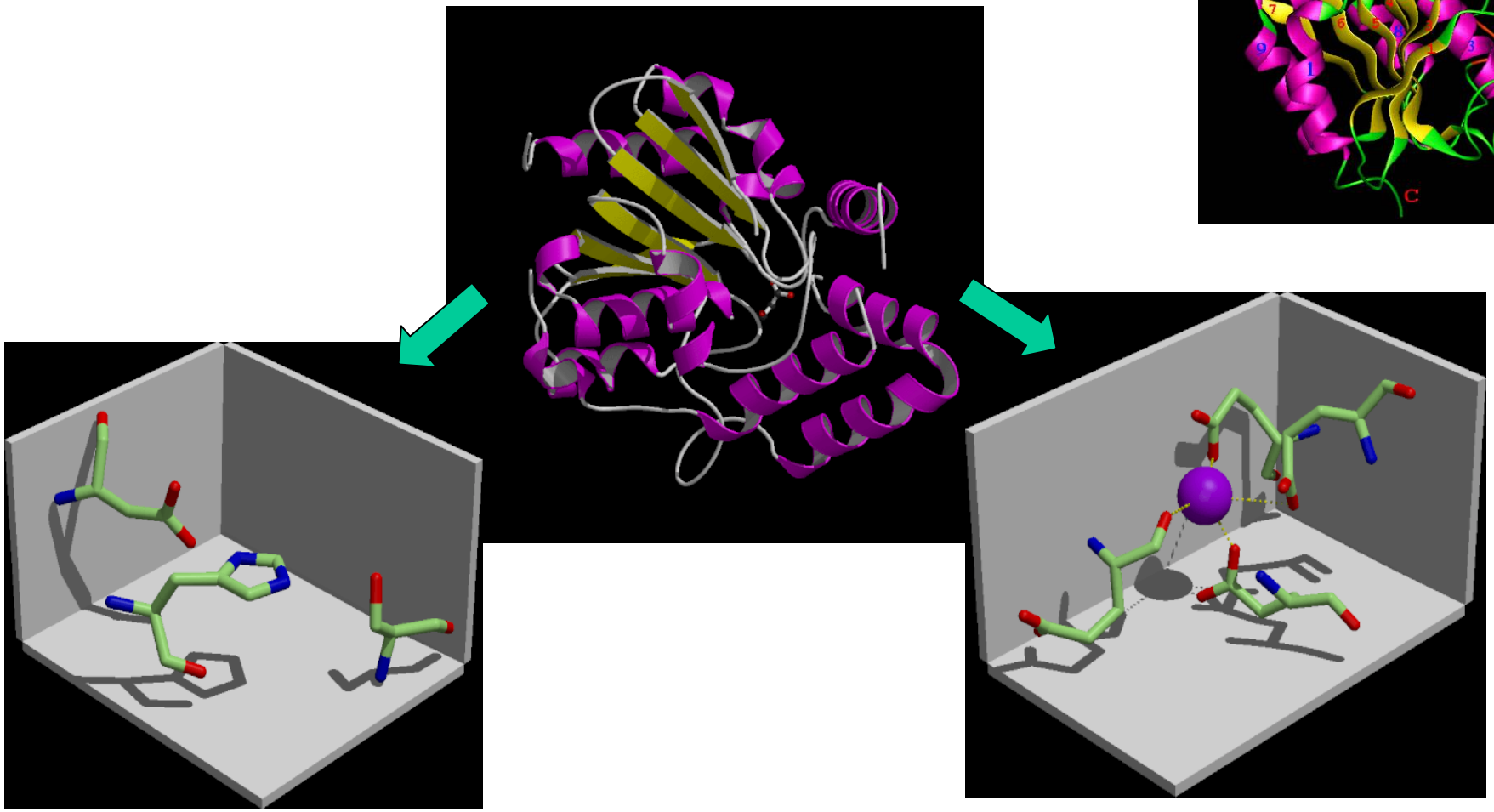
View

Alignment



Predicting Function from Structure





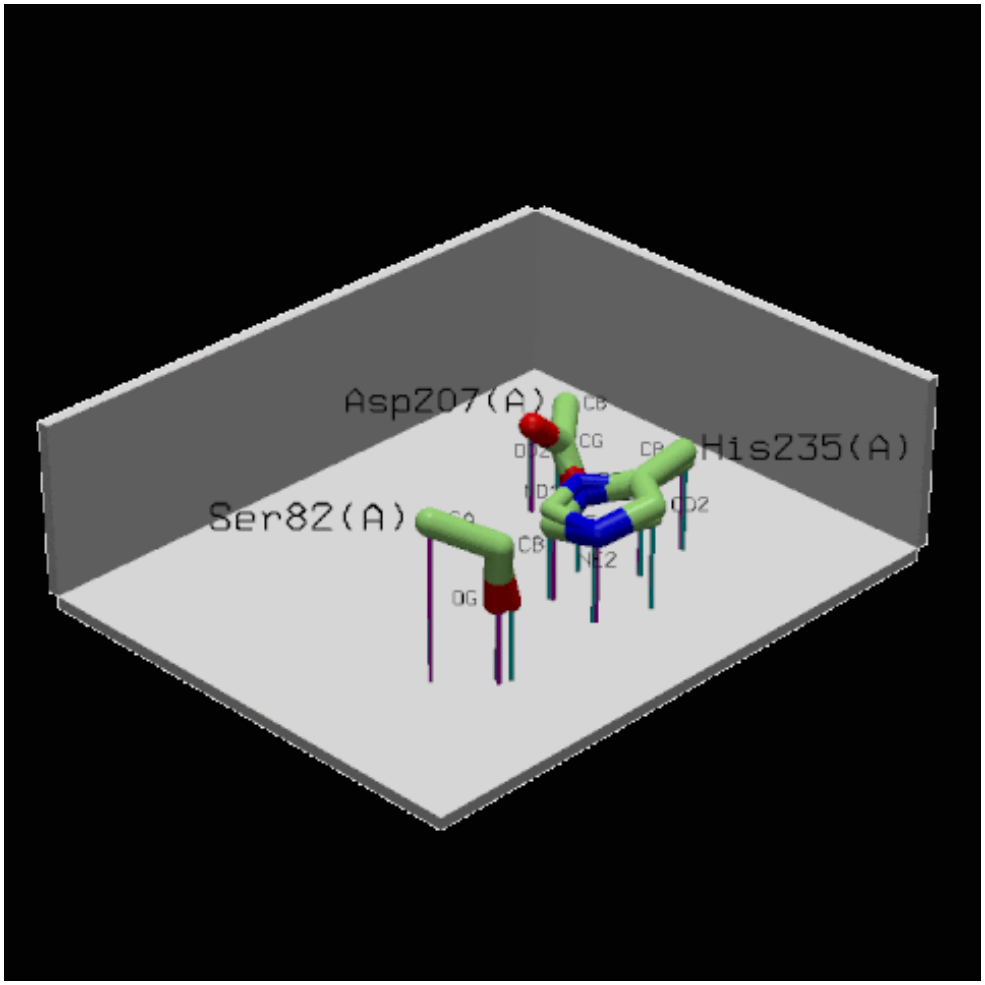
Catalytic residues

(189 active site templates)

Metal-binding sites

(~600 templates)

Proof of 3D Template Concept: BioH

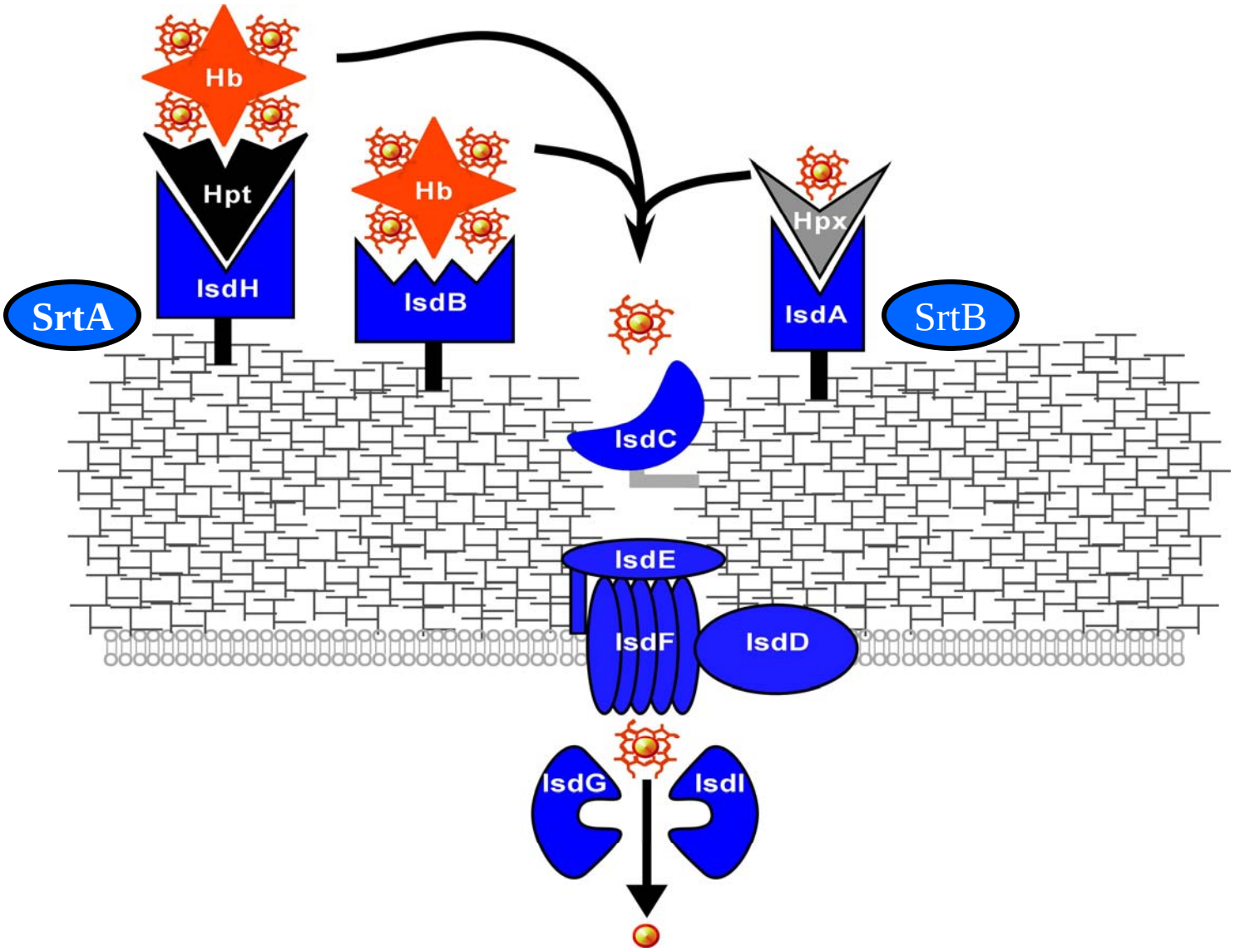


- **BioH – protein in *E. coli* involved in biotin metabolism – function unknown**
- **Crystal structure determined**
- **Catalytic triad identified**
- **Now proven experimentally to be correct**
- **BioH is carboxyesterase**

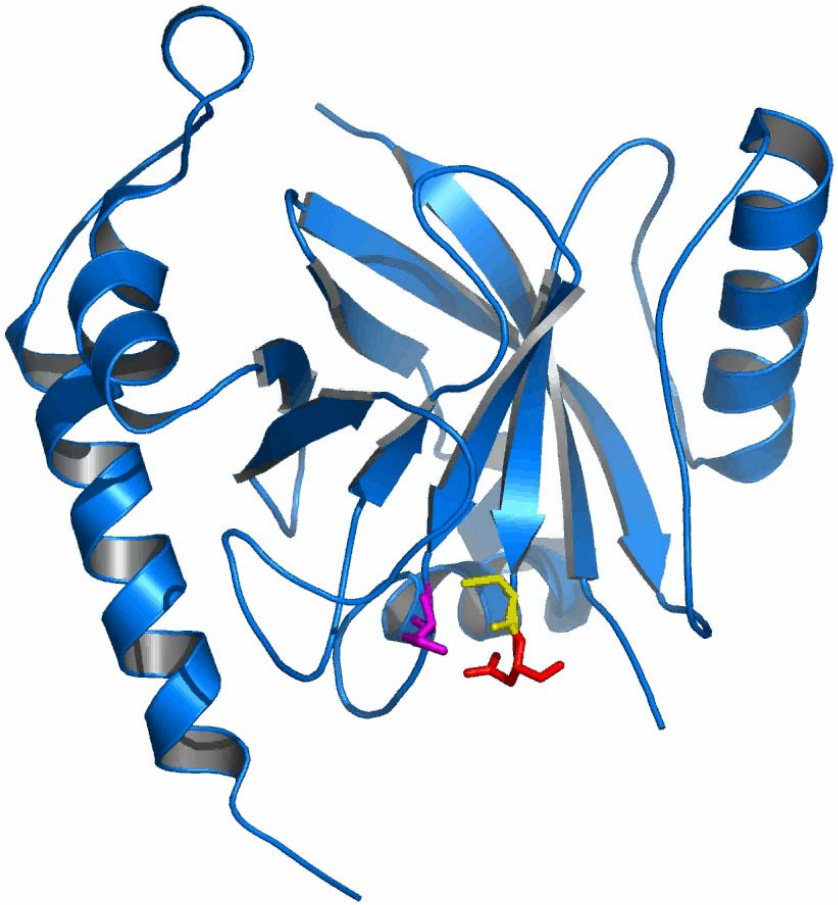
Template shown in thin bonds,
actual side chains in thick

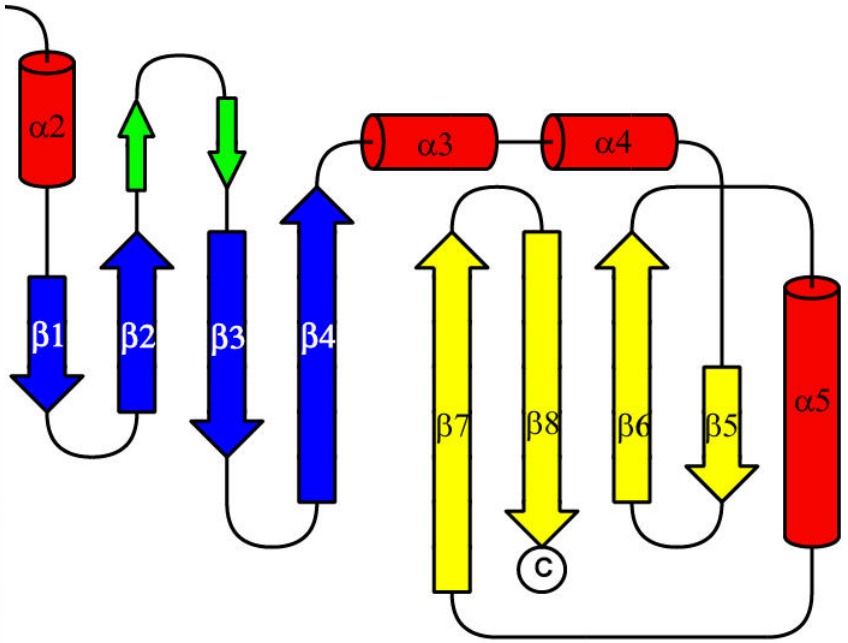
- **Ser-His-Asp catalytic triad (rmsd = 0.26Å)**
- **All 3 residues highly conserved**
- **At site of ligand binding**

A Model for Isd-mediated Heme-iron Transport in Gram Positive Pathogens



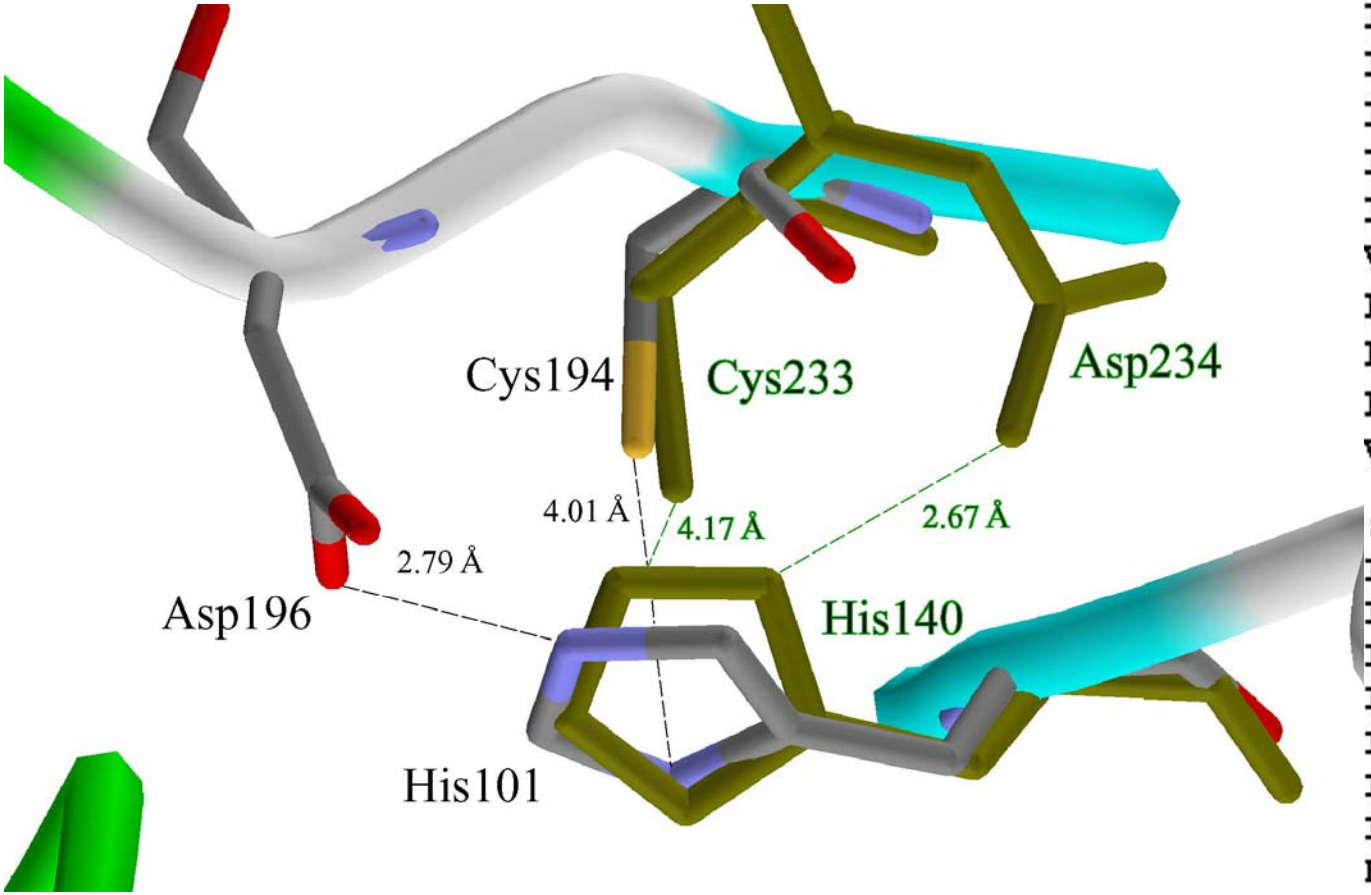
Structure of B. anthracis *Sortase B is Similar to Sortase B from* S. aureus





Active site contains
Catalytic triad Cys-His-Asp

Active Site of Sortases B from S. aureus *and* B. anthracis *Shows Catalytic Trial (with a Caveat)*




```
:::*.*.: .. **  :
ILYGHRMKDGSMFGSL  Ba
ILYGHHVGDNTMFDVL  Sa
ILYGHRMKDGSMFGSL
ILYGHRMRDGSMFAQL
IIYGHNMKDGSMFADL
IIYGHNMKDGSMFADL
VVYGHNMRNDTMFAQI
LIYGHHMAGNAMFGEI
LIYGHHMAGNAMFGEI
LIYAHHMAGNVMFGEL
VIYGHHIKGGKMFGAL

::::  **         *:::
IVTLSTCDYALDPEAGRLVV  Ba
IMTLSTCEDAYSETTKRIVV  Sa
IVTLSTCDYALDPEAGRLVV
IITLSTCDYRLDRDRGRLVV
IITLSTCDTEKDYEKGRMVI
IITLSTCDTEKDYEKGRMVI
ILTLITCGYDFVN--ARIVV
FVAFSTCENFSTDN--RVIV
FVAFSTCENFSTDN--RVIV
FVALSTCEDMTTDG--RIIV
LITLFTCEYSAQNG--RLVV
```

Impact of Structural Genomics

- **Structural genomics will map protein folding space and provide insight into relationship between amino acid sequence and 3D structure**
  - **New technology for cost-effective molecular biology and protein purification will evolve from the project**
  - **SG will enhance crystallographic capabilities by significantly reducing the time and cost required to determine protein 3D structures**
  - **Bottlenecks will be identified and practical solutions will be establish**
  - **All "easy" structures will be solved rapidly**
  - **The HTP technologies will be developed to handle challenging biological systems and will benefit biology and biotechnology**
- **Structural information will provide new insight into protein function**
- **We will gain complete structural understanding of many processes and pathways in the cell**
- **New hypothesis will be formulated**
- **New functional proteins will be created through rational design**
- **Libraries of genes, expression clones and proteins will be produced and will be available to public**

Why Neutron Crystallography?

- **The protein neutron structures can reveal number of important (and often unexpected) details:**
  - **Hydrogen atoms (~1% of the protein mass but critical for understanding protein packing, enzyme catalysis and interactions),**
  - **Protonation states and H-bonds,**
  - **Solvent structure – can distinguish water from other solvent components (Cl, Na, Mg, etc),**
- **Can define accurately atomic positions (unexpected stereochemistry, distortion of bond lengths)**
- **No radiation damage – native structure can be observed at wide range of temperatures,**
- **Hydrogen exchange with deuterium – solvent accessibility and dynamics,**
- **All these issues are critical to understand catalytic mechanism, interactions, solvent structure and potentially may have impact on structure-based drug design,**
- **Thus far low flux did not allow for large scale effort,**
- **With construction of the SNS there is an unique opportunity to build neutron crystallography beam line that can make an important contribution to biology, highly complementary to x-ray data**

MCSG and Collaborators

ANL/MCSG
J. Abdullah,
A. Binkowski,
C-s. Chang,
S. Clancy,
F. Collart,
M. Cuff,
I. Dementieva,
M. Donnelly,
C. Giometti,
C. Hatzos,
D. Holzle,
G. Joachimiak,
L. Lezondra,
S-s. Li,
H. Li,
Y-c. Kim,
N. Maltsev,
L. Marland,
S. Moy,
B. Nocek,
J. Osipiuk,
P. Quartey,
M. Schiffer,
L. Stols,
R-y. Wu,
Q-z. Zhang,
R-g. Zhang,
M. Zhou,

Univ. of Virginia
W. Minor,
M. Chruszcz,
M. Cyborowski,
M. Grabowski,
P. Lasota,
P. Miles,
H.-j. Quiu,
M. Zimmerman,
H. Zheng,

Northwestern Univ.
W. Anderson,
L. Shuvalova,
J. Brunzelle,
D. Miller,
G. Minasov,
S. Rajan,
X. Yang,
S. Weigand,
L. Shuvalova,
D. Still,
Y. Tang,
X. Yang,

Univ. College London and EBI UK,
J. Thornton,
C. Orengo,
M. Bashton,
R. Laskowski,
D. Lee,
R. Marsden,
D. McKenzie,
A. Todd,
J. Watson,

Univ. of Virginia
Z. Derewenda,

GE Health Systems

CNRS, Strasbourg
A. Podjarny,
A. Mitschler,
T. Petrova,

Univ. of Toronto
A. Edwards,
C. Arrowsmith,
E. Evdokimova,
J. Guthrie,
A. Khachatryan,
M. Kudrytska,
A. Savchenko,
T. Skarina,
X.(Linda) Xu,

ANL/SBC
R. Alkire,
N. Duke,
S. Ginell,
K. Lazarski,
M. Molitski,
F. Rotella,

ANL/GMCA
N. Sanishvili,

ANL/APS
D. Shu,
C. Preissner,
D. Nocher,
Y. Han,
P. Lee,
Z. Cai,

Univ. of Texas SWMC
Z. Otwinowski,
D. Borek,
A. Kudlicki,
A. Q. Mei,
M. Rowicka,
J. Zelinka,

Univ. of Chicago
O. Schneewind,
D. Missiakas,
P. Gornicki,
E. Skaar,
T. Kossiakoff,
V. Tereshko,
Q. Yang,

Funding: NIH and DOE