Crystallographic raw data: our plans and implementations within the NIH's Big Data to Knowledge resource

Wladek Minor

Rovinj, August 2015

Integrated Resource for Reproducibility in Macromolecular Crystallography

Goals:

1. Develop tools for automatically extracting and curating diffraction images and associated metadata, as well as producing detailed descriptions of all data needed for later reprocessing of the diffraction data as methods for structure determination improve.

2. Create a web-based system for semantic searching, analysis, and data mining of appropriate subsets of diffraction images and associated metadata.

- 3. Develop tools to automatically validate, preprocess, and score diffraction images, and to detect potential issues and errors.

- 4. Creation of a repository for diffraction data that did not yield an X-ray structure with the currently available methods.

- 5. Set up a pilot resource incorporating the tools developed in Aims 1-4 to collect a test set of data for development of tools and algorithms for validation and error detection.

Integrated Resource for Reproducibility in Macromolecular Crystallography

Goal:

Protein Crystallography (Structural Biology) with Speed and Finesse

100 years and still going strong

High throughput SB

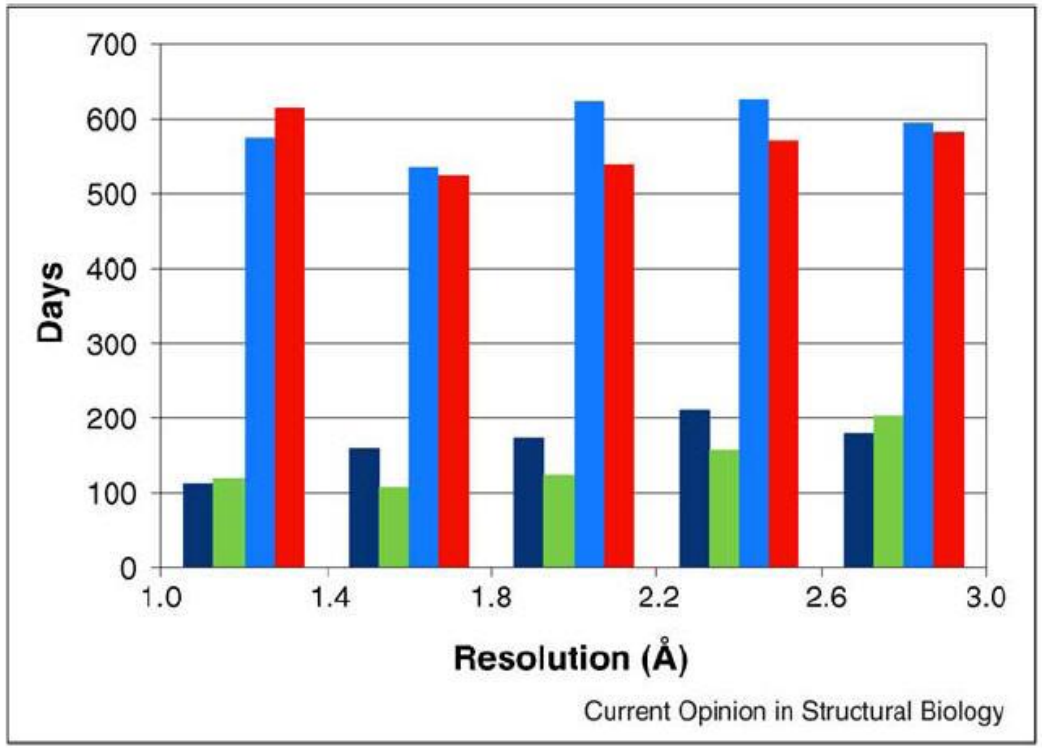
- Automatic cloning
- HT automatic expression
- HT automatic purification
- HT automatic crystallization
- HT automatic data collection
- HT automatic structure solution/refinement

High throughput SB

- Automatic cloning
- HT automatic expression
- HT automatic purification
- HT automatic crystallization
- HT automatic data collection
- HT automatic structure solution/refinement

- **Automatic paper writing**

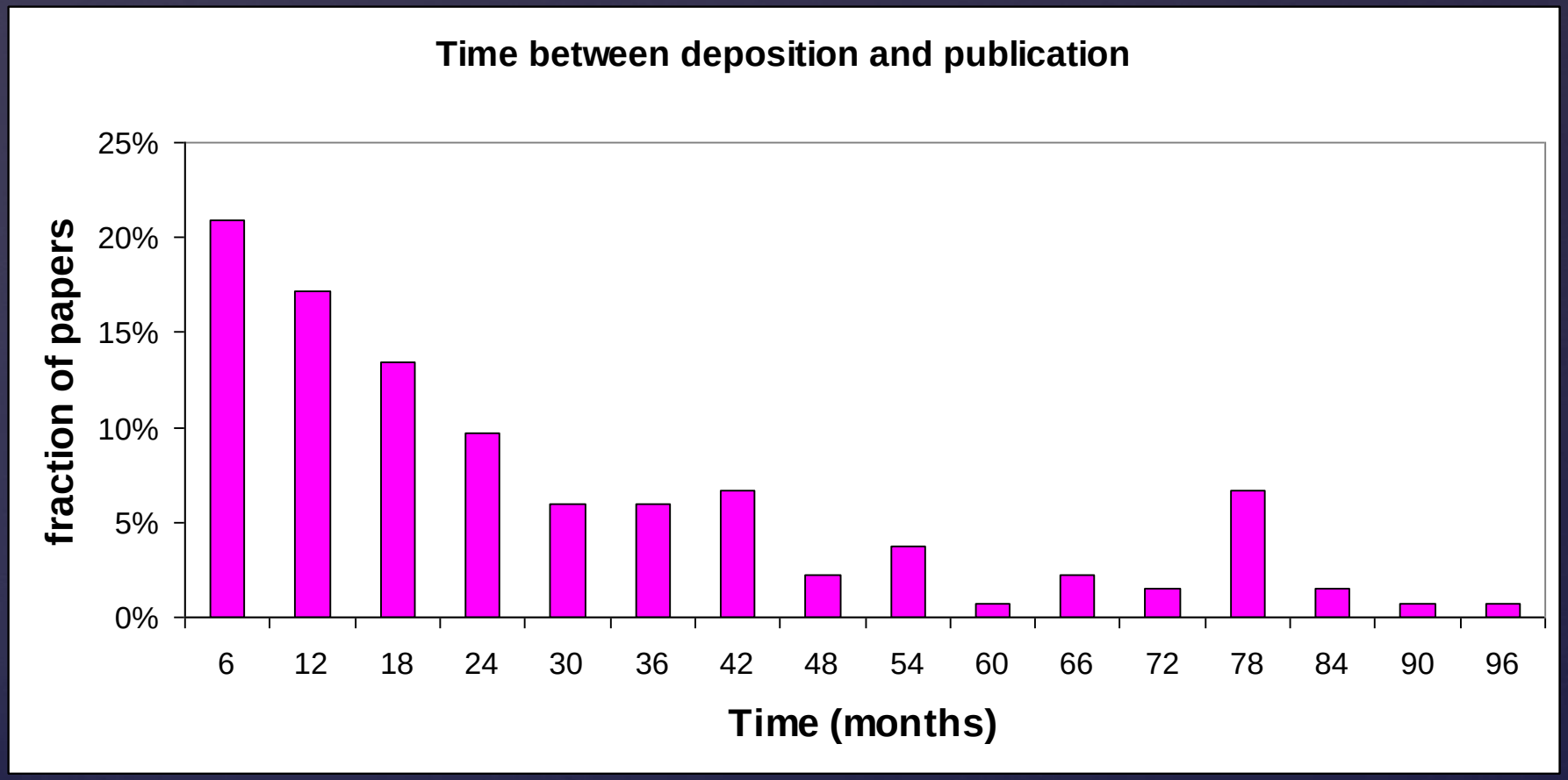
Data collection -> deposition



Average time (in days) between data collection and deposition for SG and non-SG structures. Dark blue and green bars represent SG structures, whereas light blue and red bars represent non-SG structures deposited in 2000–2004 and 2005–2009, respectively. Structures were binned by reported resolution limit (0.4 Å bin width).

Curr. Opinion in Struct. Biology (2010) **20**: 587-597

Deposition -> Publication



Curr. Opinion in Struct. Biology (2010) **20**: 587-597

Bottleneck:

Data is not information,

information is not knowledge,

knowledge is not understanding,

understanding is not wisdom

Bottleneck: brain engagement

Data is not information,

information is not knowledge,

knowledge is not understanding,

understanding is not wisdom

Clifford Stoll

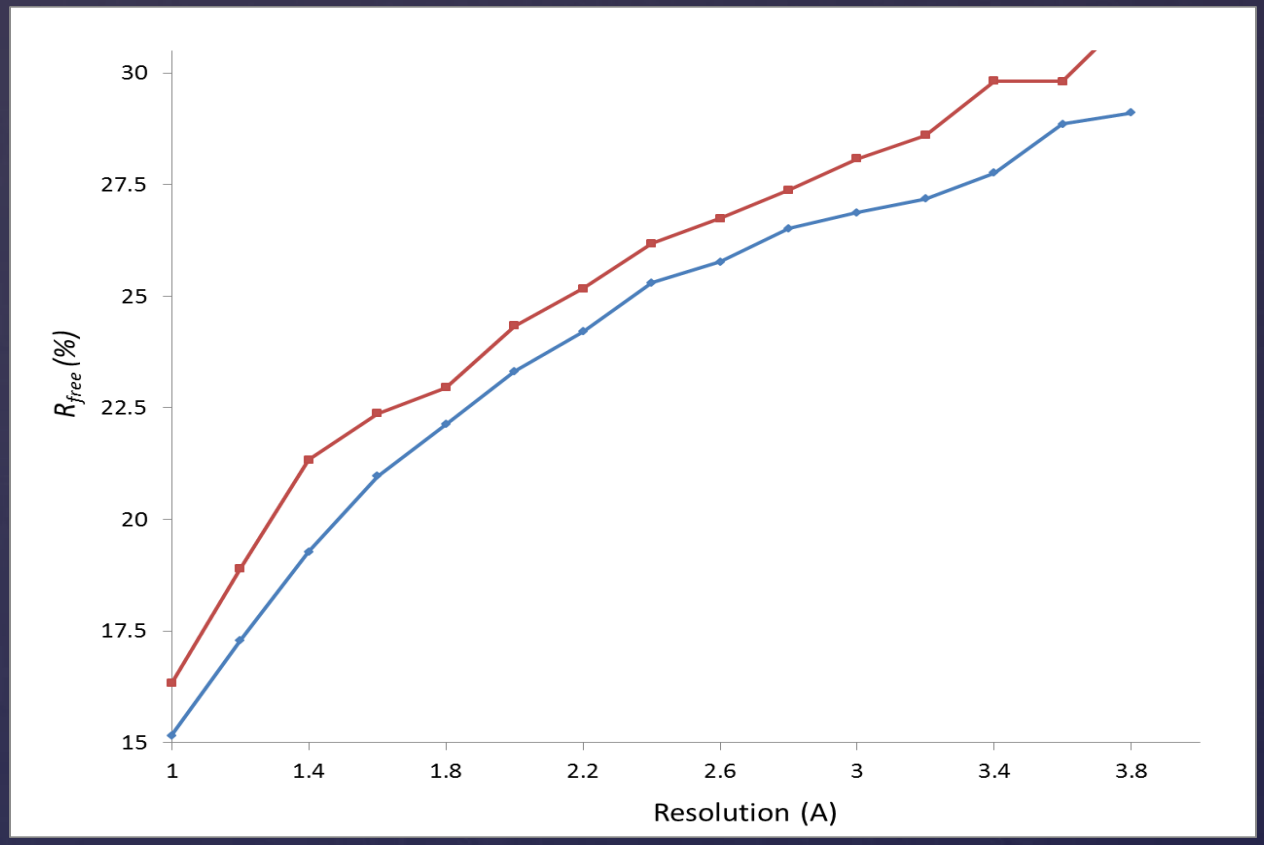
What experimenters know about data collection ?

```
REMARK   3  ESTIMATED OVERALL COORDINATE ERROR.
REMARK   3   ESU BASED ON R VALUE                              (A): NULL
REMARK   3   ESU BASED ON FREE R VALUE                         (A): NULL
REMARK   3   ESU BASED ON MAXIMUM LIKELIHOOD                   (A): NULL
REMARK   3   ESU FOR B VALUES BASED ON MAXIMUM LIKELIHOOD (A**2): NULL
REMARK   3
REMARK   3  RMS DEVIATIONS FROM IDEAL VALUES.
REMARK   3   DISTANCE RESTRAINTS.                      RMS    SIGMA
REMARK   3    BOND LENGTH                       (A) : NULL  ; NULL
REMARK   3    ANGLE DISTANCE                    (A) : NULL  ; NULL
REMARK   3    INTRAPLANAR 1-4 DISTANCE          (A) : NULL  ; NULL
REMARK   3    H-BOND OR METAL COORDINATION      (A) : NULL  ; NULL
REMARK   3
REMARK   3   PLANE RESTRAINT                    (A) : NULL  ; NULL
REMARK   3   CHIRAL-CENTER RESTRAINT        (A**3) : NULL  ; NULL
REMARK   3
REMARK   3   NON-BONDED CONTACT RESTRAINTS.
REMARK   3    SINGLE TORSION                    (A) : NULL  ; NULL
REMARK   3    MULTIPLE TORSION                  (A) : NULL  ; NULL
REMARK   3    H-BOND (X...Y)                    (A) : NULL  ; NULL
REMARK   3    H-BOND (X-H...Y)                  (A) : NULL  ; NULL
REMARK   3
REMARK   3   CONFORMATIONAL TORSION ANGLE RESTRAINTS.
REMARK   3    SPECIFIED                  (DEGREES) : NULL  ; NULL
REMARK   3    PLANAR                     (DEGREES) : NULL  ; NULL
REMARK   3    STAGGERED                  (DEGREES) : NULL  ; NULL
REMARK   3    TRANSVERSE                 (DEGREES) : NULL  ; NULL
REMARK   3
REMARK   3  ISOTROPIC THERMAL FACTOR RESTRAINTS.    RMS    SIGMA
REMARK   3   MAIN-CHAIN BOND              (A**2) : NULL  ; NULL
REMARK   3   MAIN-CHAIN ANGLE             (A**2) : NULL  ; NULL
REMARK   3   SIDE-CHAIN BOND              (A**2) : NULL  ; NULL
```

What experimenters know about data collection ?

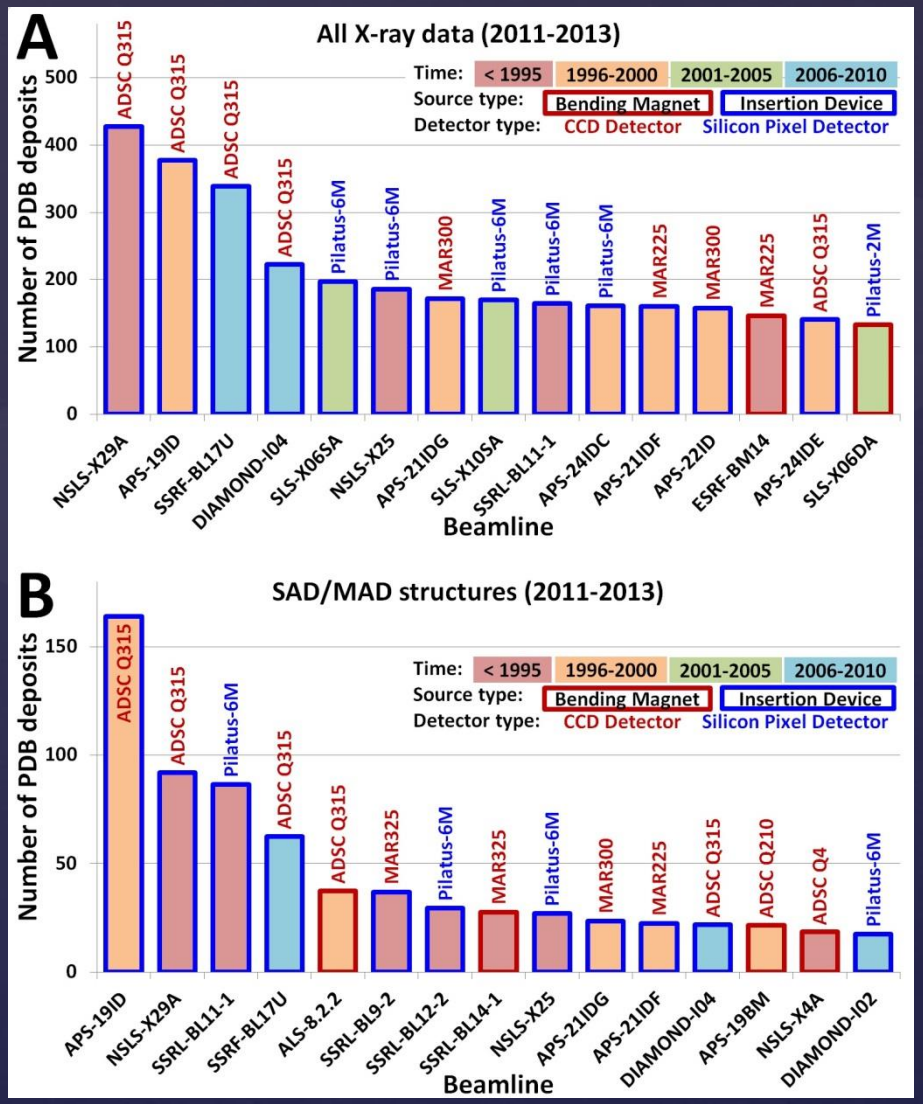
```
REMARK 200  DETECTOR TYPE                    : CCD
REMARK 200  DETECTOR MANUFACTURER            : ADSC QUANTUM 4
REMARK 200  INTENSITY-INTEGRATION SOFTWARE : BLU-ICE
REMARK 200  DATA SCALING SOFTWARE            : MOSFLM, CCP4, SCALEPACK
REMARK 200
REMARK 200  NUMBER OF UNIQUE REFLECTIONS    : 17575
REMARK 200  RESOLUTION RANGE HIGH       (A) : 2.950
REMARK 200  RESOLUTION RANGE LOW        (A) : 47.870
REMARK 200  REJECTION CRITERIA  (SIGMA(I)) : NULL
REMARK 200
REMARK 200 OVERALL.
REMARK 200  COMPLETENESS FOR RANGE      (%) : 94.3
REMARK 200  DATA REDUNDANCY                 : 3.200
REMARK 200  R MERGE                     (I) : 0.07800
REMARK 200  R SYM                       (I) : 0.08600
REMARK 200  <I/SIGMA(I)> FOR THE DATA SET   : NULL
REMARK 200
REMARK 200 IN THE HIGHEST RESOLUTION SHELL.
REMARK 200  HIGHEST RESOLUTION SHELL, RANGE HIGH (A) : 2.95
REMARK 200  HIGHEST RESOLUTION SHELL, RANGE LOW  (A) : 3.03
REMARK 200  COMPLETENESS FOR SHELL      (%) : 92.4
REMARK 200  DATA REDUNDANCY IN SHELL        : 2.80
REMARK 200  R MERGE FOR SHELL           (I) : NULL
REMARK 200  R SYM FOR SHELL             (I) : NULL
REMARK 200  <I/SIGMA(I)> FOR SHELL          : NULL
REMARK 200
REMARK 200 DIFFRACTION PROTOCOL: SINGLE WAVELENGTH
REMARK 200 METHOD USED TO DETERMINE THE STRUCTURE: MAD SE-MET
REMARK 200 SOFTWARE USED: SNB, MLPHARE, CCP4, SOLVE, HKL, RESOLE
REMARK 200 STARTING MODEL: NULL
REMARK 200
REMARK 200 REMARK: NULL
REMARK 280
REMARK 280 CRYSTAL
REMARK 280 SOLVENT CONTENT, VS   (%): NULL
REMARK 280 MATTHEWS COEFFICIENT, VM (ANGSTROMS**3/DA): NULL
REMARK 280
```

Unexpected correlation?



Average R_{free} by resolution bin (with a width of 0.2 Å for X-ray crystallography PDB structures deposited after January 1, 2001, divided into two groups by the number of missing data items ("NULLs") in the PDB file. The means for "high-completion" deposits (20 NULLs or less) are shown in blue, and the means for "low-completion" deposits (50 or more NULLs) are shown in red.

Where we should collect data ?



Zheng *et al*, *Expert Opinion on Drug Discovery(2014) 9: 125-37*

Diffraction experiment - the last experiment before deposition to PDB

Dataset – 2minutes, sample change 2minutes -> 10minutes

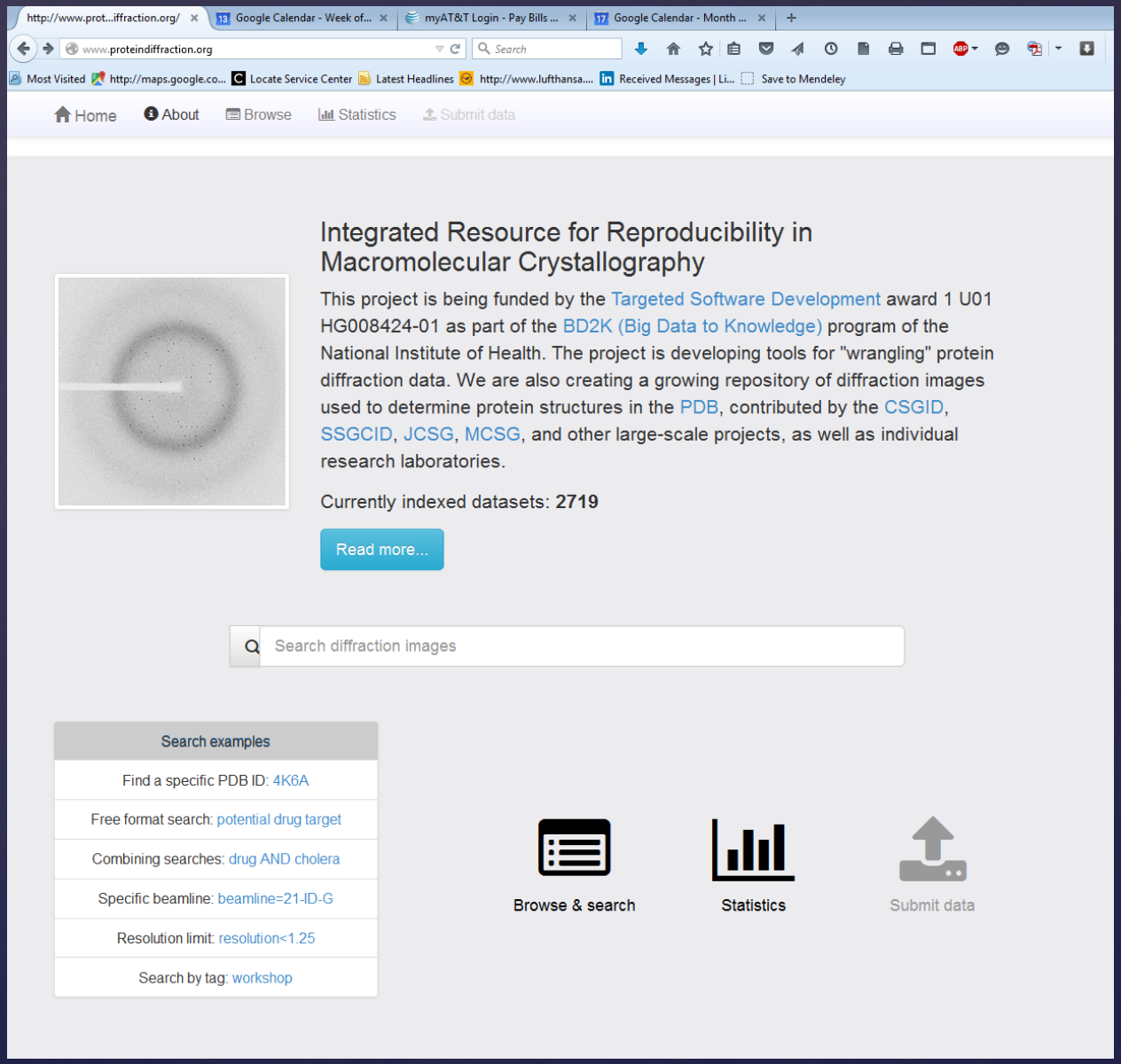
6 datasets/hour -> 144 datasets/day

180 days -> 25920 datasets/day -> 2.5 PDB

125 synchrotron stations -> 324 PDB

Efficiency -> 0.3%

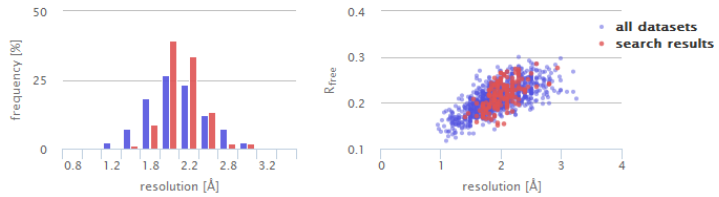
http://www.proteindiffraction.org



http://www.proteindiffraction.org

Home | About | Browse | Statistics | Submit data | beamline=home_source

There were **170** results found for this search, out of a total 2719 diffraction experiments.

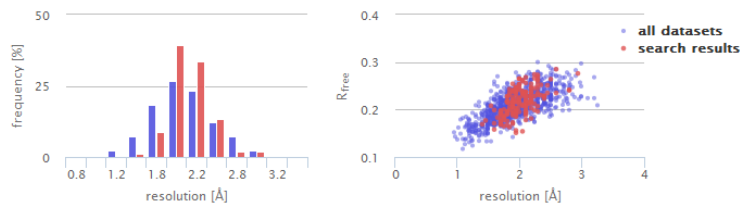


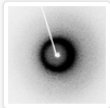
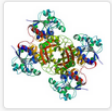
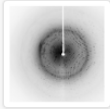
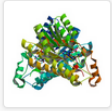
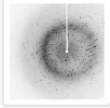
Project	Structure	Resolution	Beamline	Date
3DR6 (CSGID)	Ynca, a putative ACETYLTRANSFERASE from Salmonella typhimurium	1.75 Å	Home source	
3LNT (SSGCID)	Phosphoglyceromutase from Burkholderia Pseudomallei 1710B with bound maloni...	2.10 Å	Home source	
3DR3 (CSGID)	Idp00107, a potential N-acetyl-gamma-glutamylphosphate reductase from Shige...	2.00 Å	Home source	
3UJH (SSGCID)	Substrate-bound Glucose-6-phosphate isomerase from Toxoplasma gondii	2.10 Å	Home source	
4RGB (SSGCID)	A putative carveol dehydrogenase from Mycobacterium avium bound to NAD	1.95 Å	Home source	
3S99 (SSGCID)	A basic membrane lipoprotein from brucella melitensis, iodide soak	2.05 Å	Home source	
4XGI (SSGCID)	Glutamate dehydrogenase from Burkholderia thailandensis	2.00 Å	Home source	
3P10 (SSGCID)	2-c-methyl-d-erythritol 2,4-cyclodiphosphate synthase from Burkholderia pse...	1.70 Å	Home source	
3O2E (SSGCID)	A bol-like protein from babesia bovis	1.95 Å	Home source	
3LUZ (SSGCID)	Extragenic suppressor protein suhB from Bartonella henselae, via combined i...	2.05 Å	Home source	
3N5O (SSGCID)	Putative glutathione transferase from Coccidioides immitis bound to glutath...	1.85 Å	Home source	
4DDD (SSGCID)	An immunogenic protein from ehrlichia chaffeensis	1.90 Å	Home source	
3KM3 (SSGCID)	Eoxycytidine triphosphate deaminase from anaplasma phagocytophilum at 2.1A ...	2.10 Å	Home source	
4DDO (SSGCID)	3-oxoacyl-[acyl-carrier-protein] synthase ii from burkholderia vietnamiensi...	1.90 Å	Home source	
3S6L (SSGCID)	A YadA-like head domain of the trimeric autotransporter adhesin BoaA from B...	2.30 Å	Home source	
3S6M (SSGCID)	The structure of a Peptidyl-prolyl cis-trans isomerase from Burkholderia ps...	1.65 Å	Home source	

http://www.proteindiffraction.org

Home | About | Browse | Statistics | Submit data | beamline=home_source

There were **170** results found for this search, out of a total 2719 diffraction experiments.



Project	Structure	Resolution	Beamline	Date
3DR6 (CSGID)	Ynca, a putative ACETYLTRANSFERASE from Salmonella typhimurium	1.75 Å	Home source	
3LNT (SSGCID)	Phosphoglyceromutase from Burkholderia Pseudomallei 1710B with bound maloni...	2.10 Å	Home source	
3DR3 (CSGID)	Idp00107, a potential N-acetyl-gamma-glutamylphosphate reductase from Shige...	2.00 Å	Home source	
 	First author: A.U. Singer Uniprot: P59310 Gene name: argC	Space group: P 62 2 2 R/R_{free}: 0.17/0.22 I/σ in last shell: 10.1	View dataset details Download all images (1.1GB) CSGID website for IDP00107 PDB website for 3DR3	
3UJH (SSGCID)	Substrate-bound Glucose-6-phosphate isomerase from Toxoplasma gondii	2.10 Å	Home source	
4RGB (SSGCID)	A putative carveol dehydrogenase from Mycobacterium avium bound to NAD	1.95 Å	Home source	
 	First author: T.E. Edwards Uniprot: A0QDP5 Gene name: not available	Space group: C 2 2 21 R/R_{free}: 0.21/0.24 I/σ in last shell: 2.8	View dataset details Download all images (1.0GB) SSGCID website for SSGCID-MyavA.01326.g PDB website for 4RGB	
3S99 (SSGCID)	A basic membrane lipoprotein from brucella melitensis, iodide soak	2.05 Å	Home source	
4XGI (SSGCID)	Glutamate dehydrogenase from Burkholderia thailandensis	2.00 Å	Home source	
3P10 (SSGCID)	2-c-methyl-d-erythritol 2,4-cyclodiphosphate synthase from Burkholderia pse...	1.70 Å	Home source	
 	First author: L. Baugh Uniprot: Q3JRA0 Gene name: not available	Space group: C 1 2 1 R/R_{free}: 0.16/0.19 I/σ in last shell: 5.7	View dataset details Download all images (0.3GB) SSGCID website for BupsA.00122.a PDB website for 3P10	
3O2E (SSGCID)	A bol-like protein from babesia bovis	1.95 Å	Home source	

Target status and path to success

Space Tree - Mozilla Firefox

File Edit View History Bookmarks Tools Help

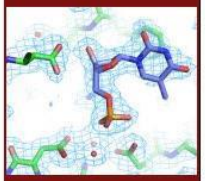
http://csgid.org/csgid/cake/space_tree/view/IDP00044

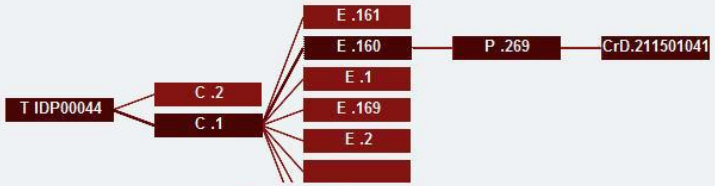
Most Visited | Getting Started | Latest Headlines | http://www.lufthansa...

Center for Structural Genomics of Infectious Diseases

Home | Target List | Selection | Community Requests | XML files | Diffraction Images | Progress | Homolog Search | Statistics | Help

- Consortium
- Overview
- Investigators
- Targets
- 3-D Structures
- Publications
- Related Sites
- Progress
- Collaborators
- Basecamp
- Log In





Expression
(simple view)

Target	IDP00044
Expression Suffix	160
Protein Clone	1434
Local Clone ID	9
Local Protein Target ID	595
Local Expression ID	160
Experiment Date Start	2008-09-29
Experiment Date End	2008-09-30
Person	Majka Klimecka
Lab	University of Virginia
Protocol ID	uva_expression_native_1
Status	success
Expression Organism	E. coli
Expression Strain	BL21 Codon Plus RILT
Media Type	LB
Media Volume	4 mL
Growth Temperature	37 °C
Induction Temperature	20 °C
Induction Reagent	0.5 mM IPTG
Experiment Type	production
Expression Level	80 %
Solubility Level	80 %

For any help please contact support@csgid.org.

CAKEPHP POWER

Metal binding site validation: *CheckMyMetal* server

ID	Res	Atom	Valence	BV symmetry	Geometry	RMSD geometry angles	Missing vertices	Bidentate	CBVS	Alternative metal
400:A	_MG	MG--	2.1	0.111	Octahedral	5.36	0	0	4.41	
400:B	_MG	MG--	2.06	0.079	Octahedral	4.13	0	0	4.32	

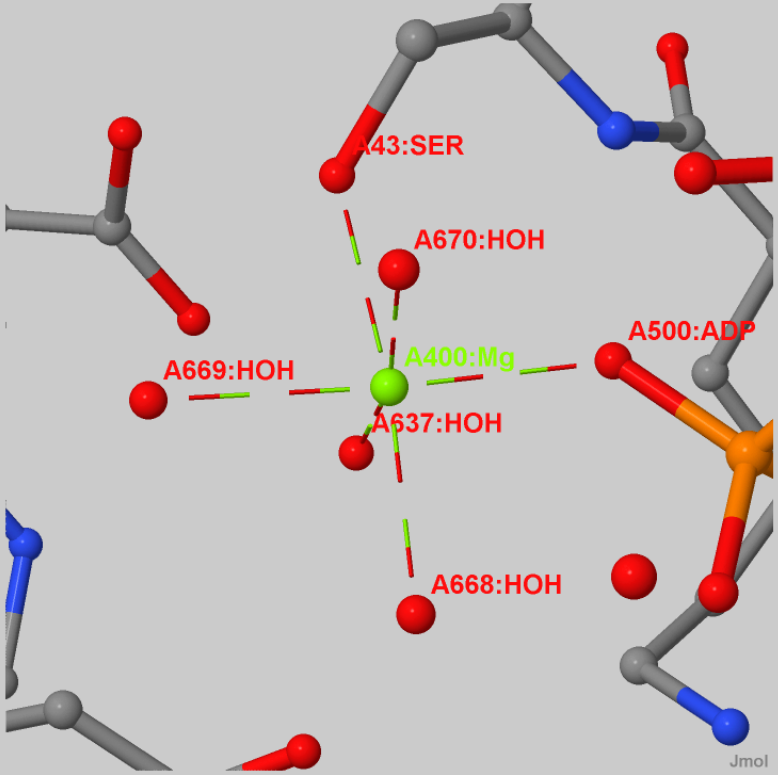
Mouse click action:

None Center Distance Label

Zoom In Zoom Out Center

Basic controls:

Left-Click to rotate
Shift-Left-Click up & down to zoom
Right-Click for Jmol's context menu



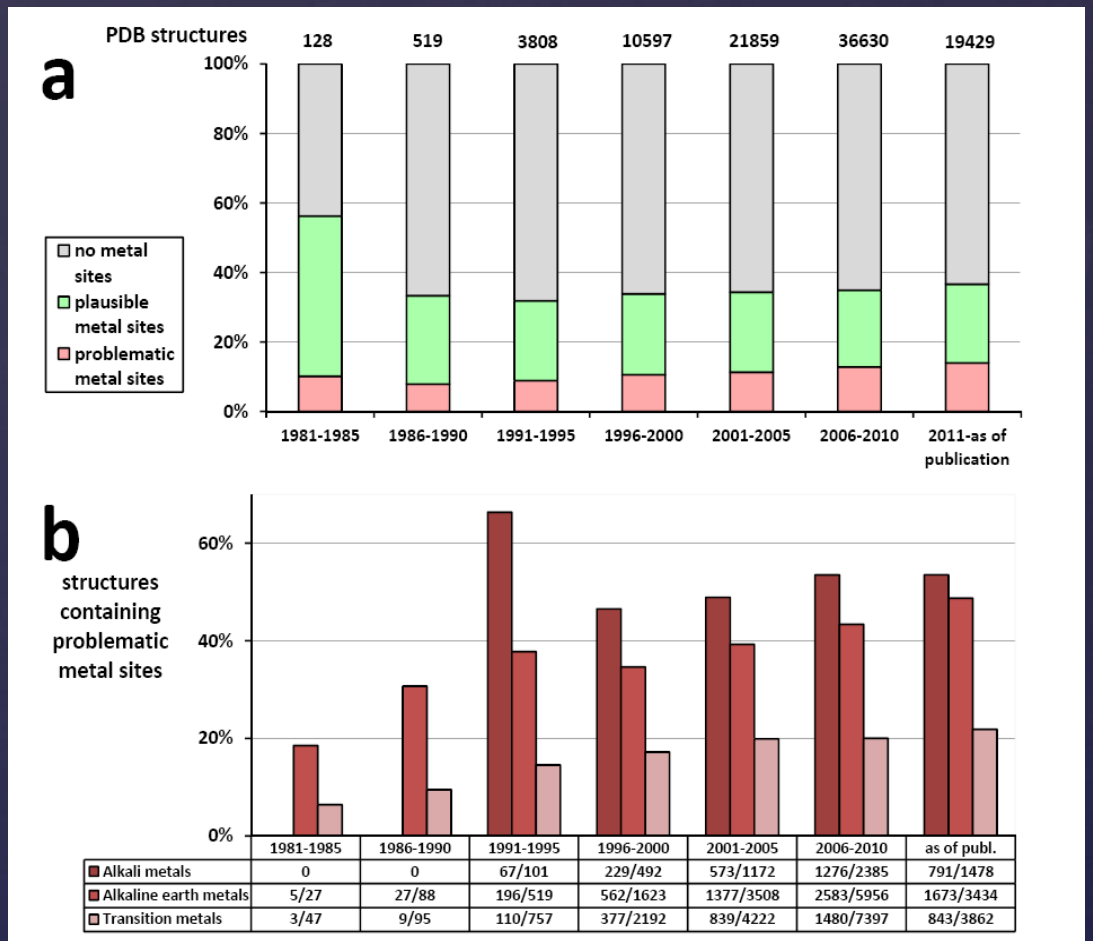
Use the buttons below to control the view

Residue Name:	On	Off
Metal Distances:	On	Off
Protein Cartoon:	On	Off
Spin:	On	Off
Antialiasing:	On	Off

Legend	Explanation
Valence	Summation of bond valence values for an ion binding site
BV symmety	Summation of bond valence vectors, weighted by bond valence values. Increase when the coordination sphere is not symmetrical due to incompleteness.
Geometry	Ion binding site geometry, as calculated by the NEIGHBORHOOD algorithm
RMSD geometry angles	R.M.S. Deviation of observed geometry angles (L-M-L angles) compared to ideal geometry, in degrees
Missing vertices	Number of sites that is not occupied in the coordination sphere for the assigned geometry
Bidentate	Number of residues that form a bidentate interaction instead of being considered as multiple ligands
CBVS	Calcium Bond Valence Sum, used for alternative metal(s) prediction [Muller,P. et al. (2003) Is the bond-valence method able to identify metal atoms in protein structures? Acta Crystallogr. D Biol. Crystallogr., 59, 32-37.]
Alternative metal	A list of alternative metal(s) is proposed in descending order of confidency, assuming metal environment is accurately determined

Zheng *et al Nature protocols* (2014) 9: 156-70

Metals in PDB



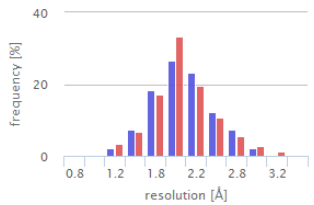
	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2010	as of publ.
Alkali metals	0	0	67/101	229/492	573/1172	1276/2385	791/1478
Alkaline earth metals	5/27	27/88	196/519	562/1623	1377/3508	2583/5956	1673/3434
Transition metals	3/47	9/95	110/757	377/2192	839/4222	1480/7397	843/3862

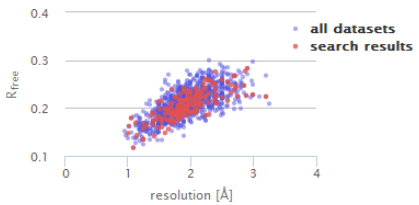
Zheng *et al Nature protocols* (2014)9: 156-70

http://www.proteindiffraction.org

Home | About | Browse | Statistics | Submit data | beamline=21-ID-G

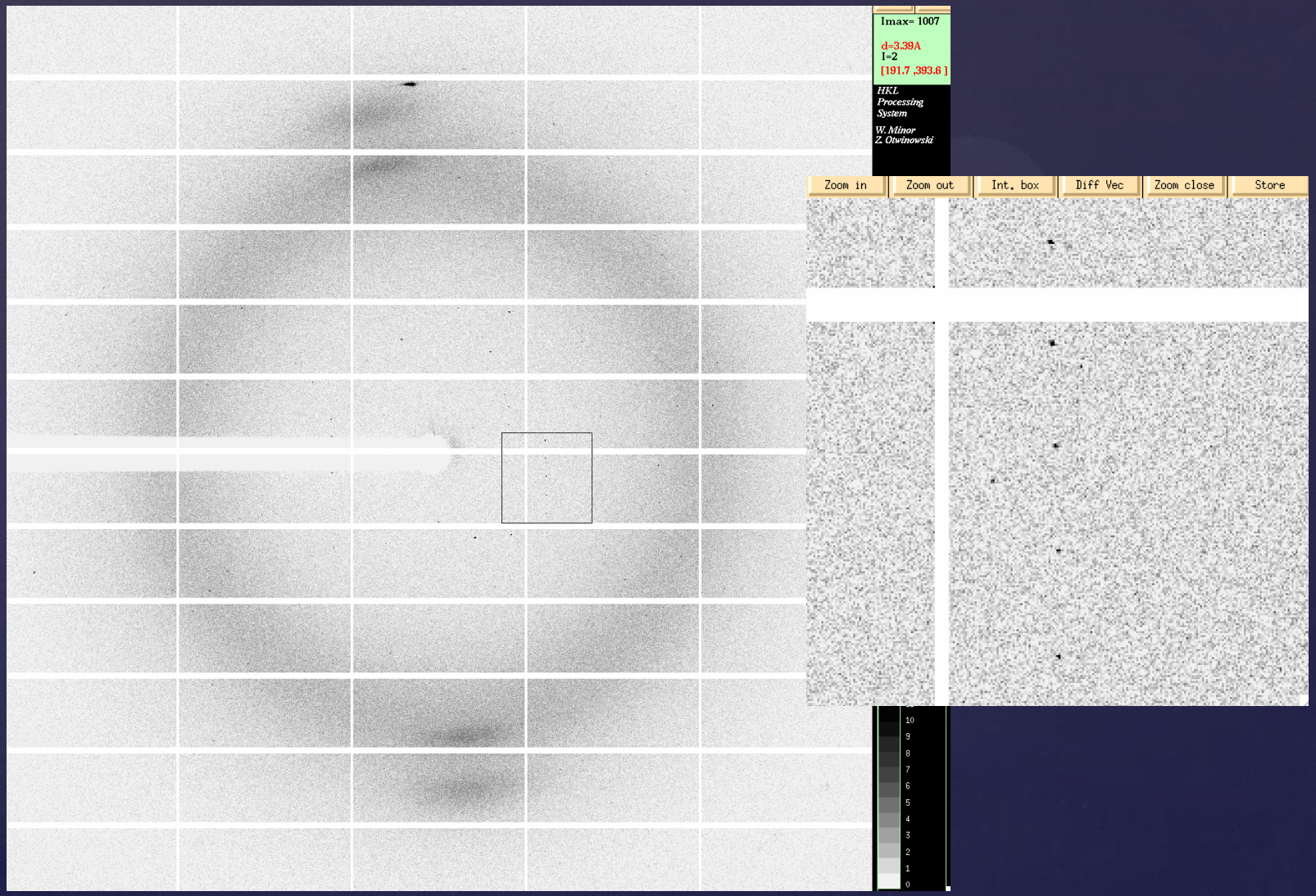
There were **180** results found for this search, out of a total 2719 diffraction experiments.



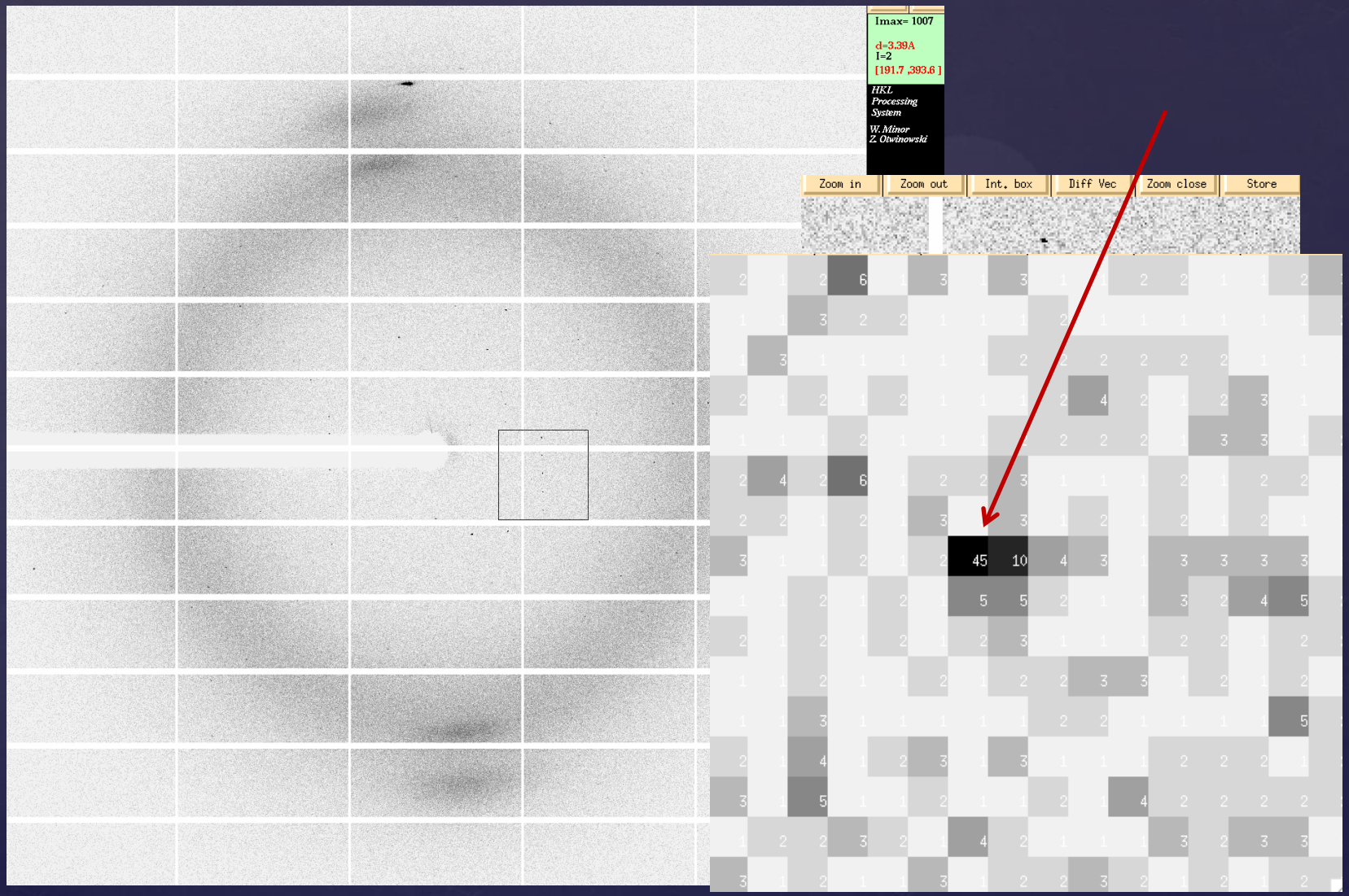


Project	Structure	Resolution	Beamline	Date
4EQ9 (CSGID)	1.4 Angstrom Crystal Structure of ABC Transporter Glutathione-Binding Prote...	1.40 Å	APS / 21-ID-G	
4X9K (CSGID)	Beta-ketoacyl-acyl carrier protein synthase III-2 (FabH2)(C113A) from Vibri...	1.61 Å	APS / 21-ID-G	
4EAQ (MCSG)	Thymidylate Kinase from Staphylococcus aureus in complex with 3'-Azido-3'-D...	1.85 Å	APS / 21-ID-G	
3QTB	Universal stress protein from Archaeoglobus fulgidus in complex with dAMP	2.10 Å	APS / 21-ID-G	
4JBE (MCSG)	1.95 Angstrom Crystal Structure of Gamma-glutamyl phosphate Reductase from ...	1.95 Å	APS / 21-ID-G	
4KWT (CSGID)	Unliganded anabolic ornithine carbamoyltransferase from Vibrio vulnificus a...	1.86 Å	APS / 21-ID-G	
4GIB (CSGID)	2.27 Angstrom Crystal Structure of beta-Phosphoglucomutase (pgmB) from Clos...	2.27 Å	APS / 21-ID-G	
4JG9 (CSGID)	X-ray Crystal Structure of a Putative Lipoprotein from Bacillus anthracis	2.42 Å	APS / 21-ID-G	
4OC9 (CSGID)	2.35 Angstrom resolution crystal structure of putative O-acetylhomoserine (...	2.35 Å	APS / 21-ID-G	
4HVN (MCSG)	Hypothetical protein with ketosteroid isomerase-like protein fold from Cate...	1.95 Å	APS / 21-ID-G	
3NNT (CSGID)	K170m Mutant of Type I 3-Dehydroquinate Dehydratase (aroD) from Salmonella ...	1.60 Å	APS / 21-ID-G	
3M07 (CSGID)	1.4 Angstrom Resolution Crystal Structure of Putative alpha Amylase from Sa...	1.40 Å	APS / 21-ID-G	
3LAY (CSGID)	Alpha-Helical barrel formed by the decamer of the zinc resistance-associate...	2.70 Å	APS / 21-ID-G	
4OJ7 (SSGCID)	Chorismate Mutase from Burkholderia thailandensis	2.15 Å	APS / 21-ID-G	
3TMQ (SSGCID)	A 2-dehydro-3-deoxyphosphooctonate aldolase from Burkholderia pseudomallei ...	2.10 Å	APS / 21-ID-G	
3IJ3 (CSGID)	1.8 Angstrom Resolution Crystal Structure of Cytosol Aminopeptidase from Co...	1.80 Å	APS / 21-ID-G	

Optimal data collection ?



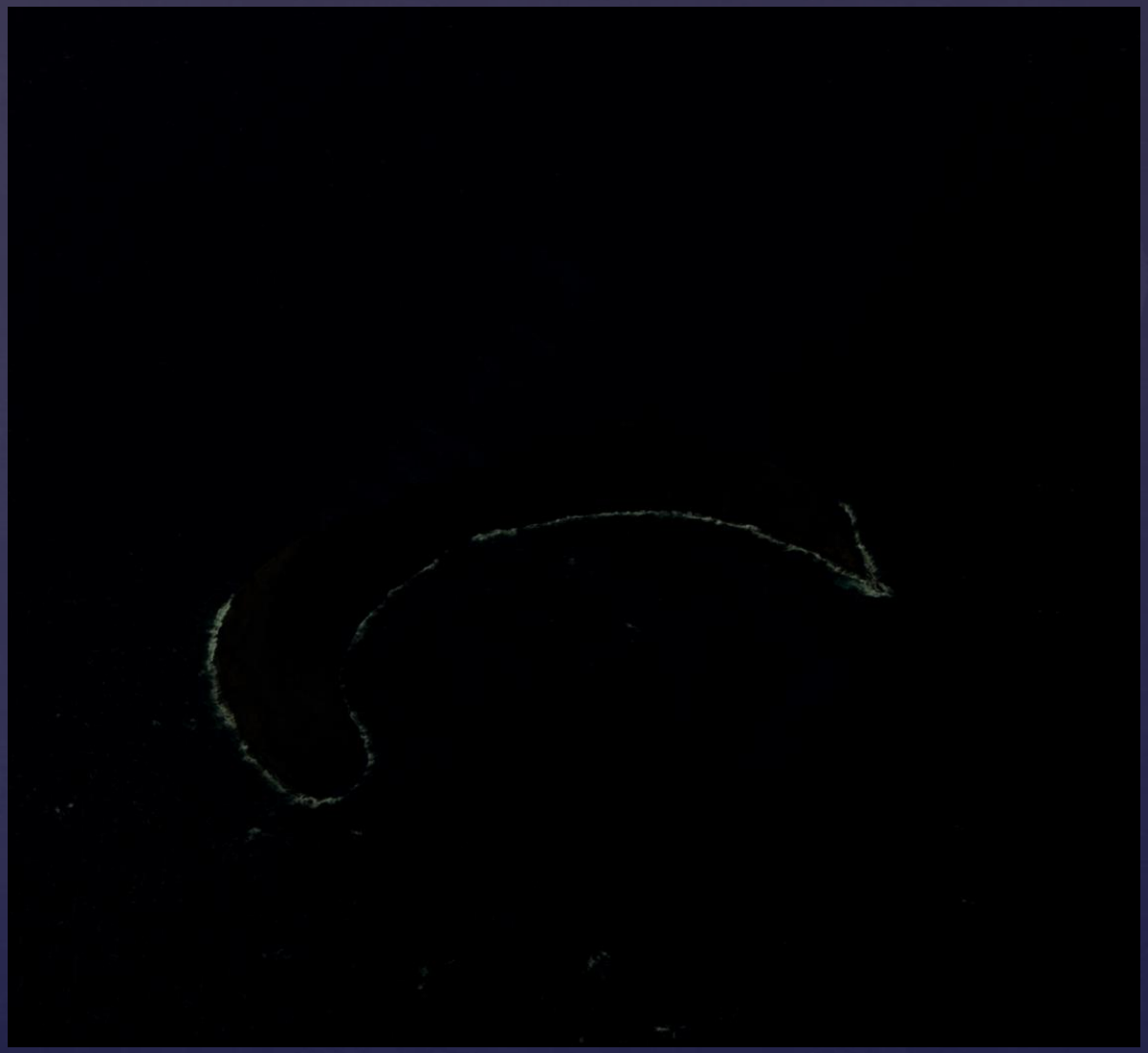
Optimal data collection ?



Header – is CBF header a MAH ?

```
# 2015/May/06 10:30:40
# Pixel_size 172e-6 m x 172e-6 m
# Silicon sensor, thickness 0.001 m
# Oscillation_axis omega
# Excluded_pixels:  badpix_mask.tif
# Chi 0.0000 deg.
# Angle_increment 0.1000 deg.
# Polarization 0.99
# file_comments
# N_oscillations 2500
# Beam_xy (1223.03, 1256.56) pixels
# Exposure_time 0.020000 s
# Phi 0.0020 deg.
# Energy_range (0, 0) eV
# Start_angle 160.6000 deg.
# Detector_distance 0.617619 m
# Detector_Voffset 0.0000 m
# Alpha 0.0000 deg.
# Flat_field: (nil)
# Threshold_setting 7619 eV
# Exposure_period 0.020950 s
# N_excluded_pixels: = 321
# Kappa 0.0020 deg.
# Tau = 0 s
```

Do you like this image?



Do you like this image ?

How expensive is bright lens ?

See more choices

Canon EF 85mm f1.2L II USM Lens for Canon DSLR Cameras - Fixed
by Canon

$1,999.00 Prime
Get it by **Monday, Aug 24**

More Buying Choices
$1,999.00 new (22 offers)
$1,499.99 used (24 offers)

Trade-in eligible for an Amazon gift card

★★★★★ ▾ 159

See Style Options

Canon EF 85mm f/1.8 USM Medium Telephoto Lens for Canon SLR Cameras - Fixed
by Canon

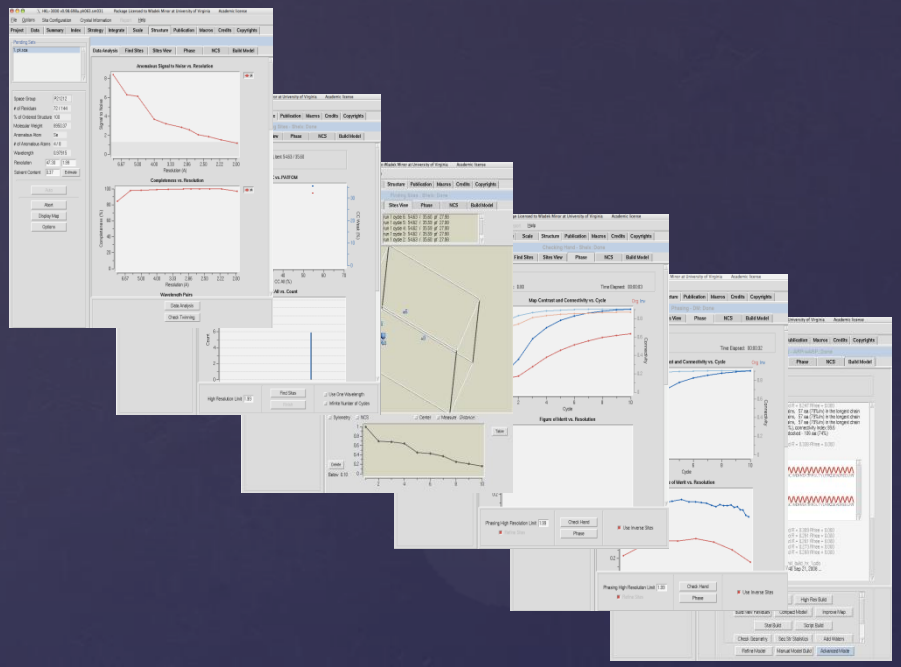
$369.00 Prime
Get it by **Monday, Aug 24**

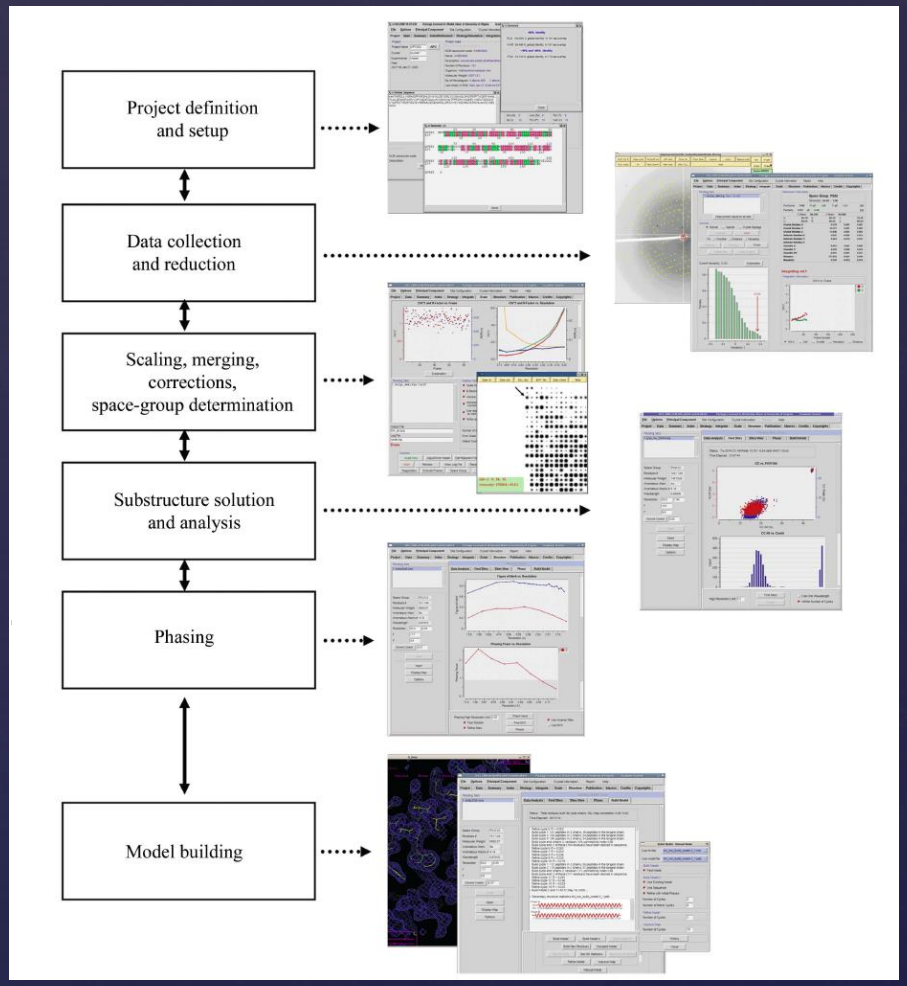
More Buying Choices
$369.00 new (27 offers)
$298.00 used (26 offers)

Trade-in eligible for an Amazon gift card

★★★★★ ▾ 770

HKL-3000 (6 mouse clicks program)



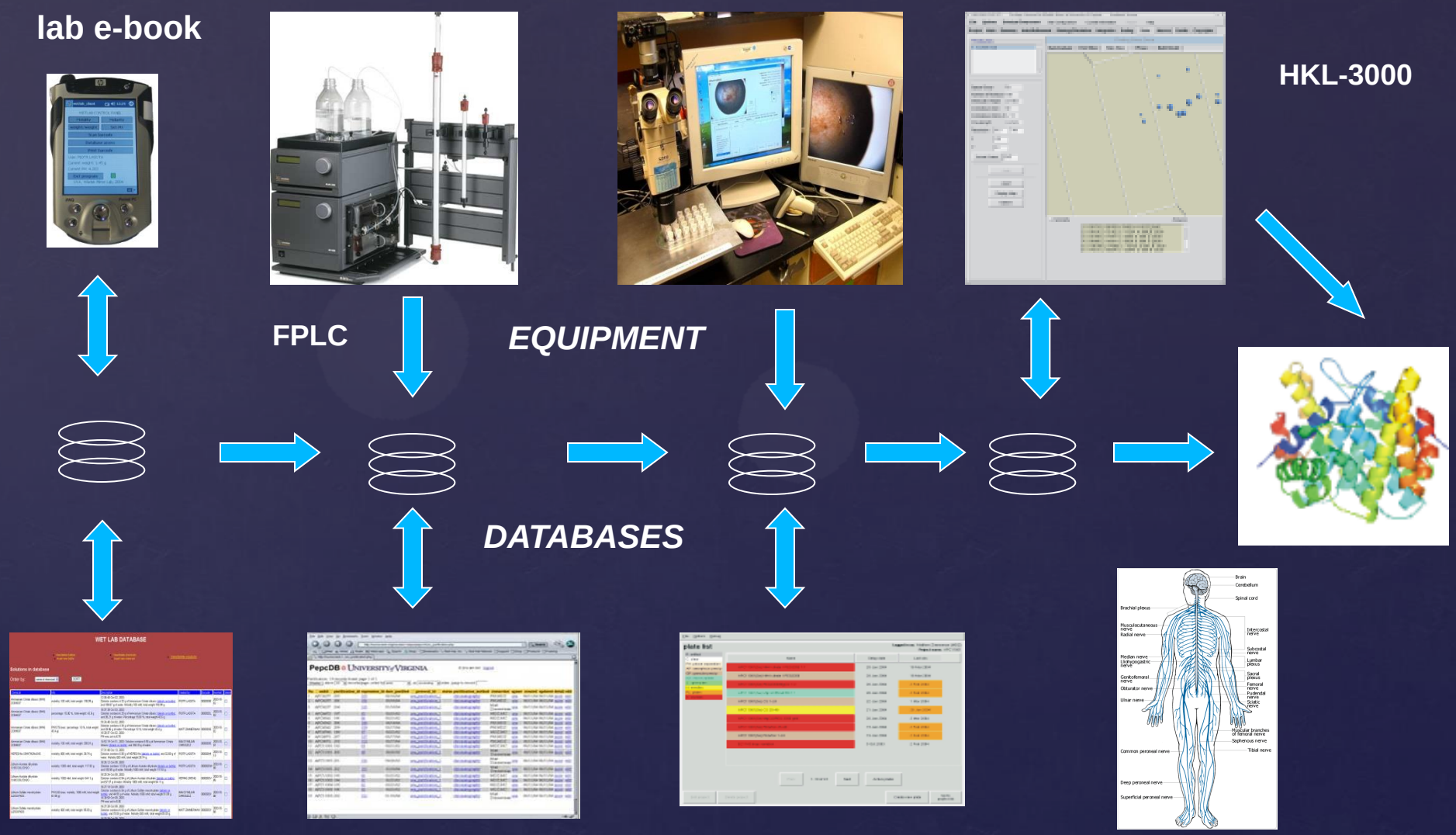


SHELXD, SHELXE
CCP4, DM, REFMAC
SOLVE, RESOLVE
ARP/WARP
O, COOT, CCP4

Acta Cryst. Section D: Biological Crystallography (2006) **62**: 859-66

HKL-3000 at SBC

Database-controlled pipeline



Big brother?

Statistics / Progress in Minor Lab LIMS by researcher

Last week (17 Apr 2015 - 24 Apr 2015)

Person	Clones	Exprs	Purifs	Macro preps	Plates	Drops	Crystals	Datasets processed	Structure refs	Kinetic assays	Thermal shift assays
Cooper, David	0	0	0	0	0	0	23	18	0	0	0
Handing, Katarzyna	0	0	0	0	0	0	51	53	13	0	0
Hou, Jing	0	0	0	0	1	30	0	1	1	0	0
Kowiel, Marcin	0	0	0	0	0	0	1	8	3	0	0
Shabalin, Ivan	0	0	0	0	0	0	125	14	9	0	0
Shumilin, Igor	0	0	0	0	0	0	0	3	2	0	0
Szlachta, Karol	0	0	0	0	0	0	34	20	3	0	0

Last month (25 Mar 2015 - 24 Apr 2015)

Acknowledgments

Wladek Minor
- Matt Zimmerman
- Marek Grabowski
- Heping Zheng
- Marcin Cymborowski
- Karol Langner (Google)
- Przemek Porebski
- Piotr Sroka
- Ivan Shabalin
- Katherine Handing

Zbyszek Otwinowski
- Dominika Borek

Andrzej Joachimiak
MCSG and SBC staff

Wayne Anderson and CSGID staff

Steve Almo and NYSGRC Staff

Ian Wilson, Marc Elsliger and JCSG staff

Steven Burley, John Westbrook and PDB staff

Tom Terwilliger
Zbyszek Dauter

Grants:
U01-HG008424
NIH GM53163, GM62414, GM74942
GM093342, GM094585, GM094662
DOE, NCI
NIAID HHSN272200700058C
NIAID HHSN272201200026C
HKL Research. Inc.