Pre-calculated protein structure alignments at the RCSB PDB site

Andreas Prlić

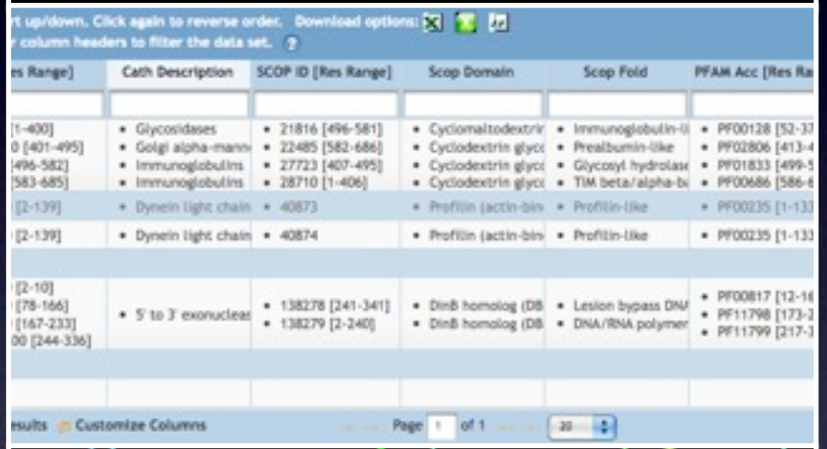
RCSB Protein Data Bank

- Archive of experimentally determined 3D structures of proteins, nucleic acids, complex assemblies

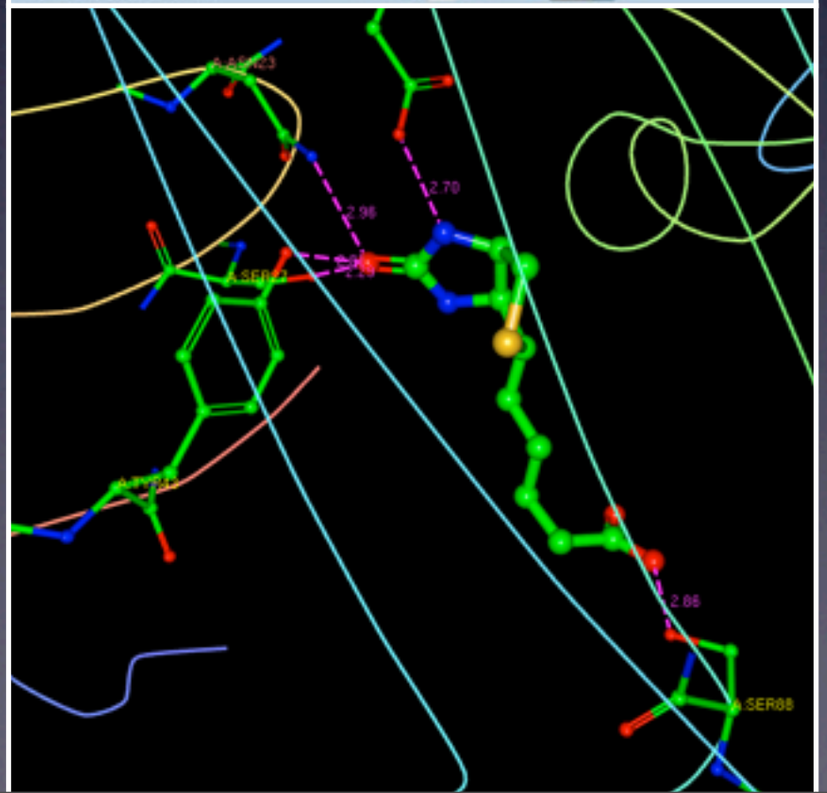
- One of the largest scientific resources in life sciences

www.pdb.org

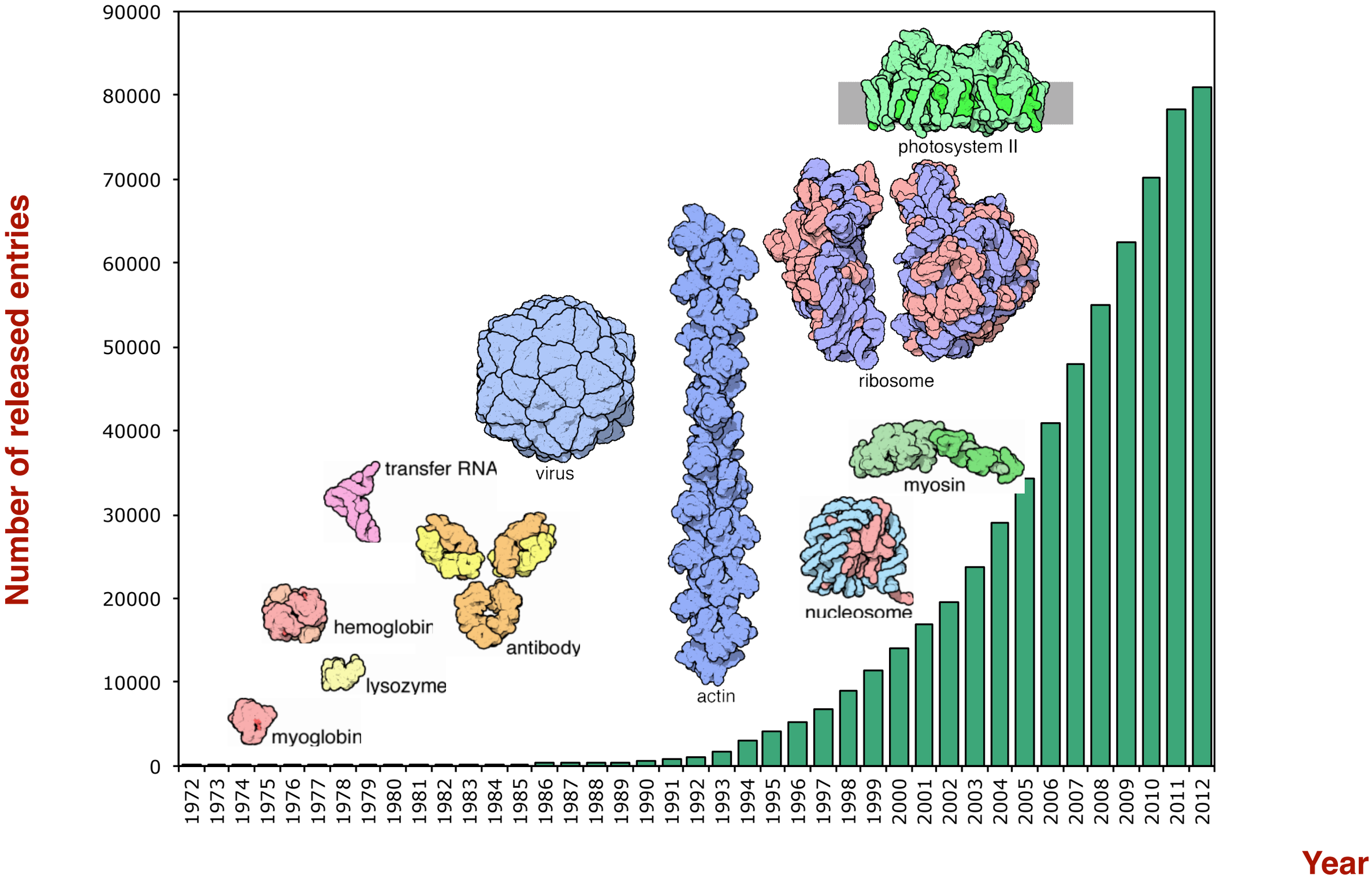
Jmol

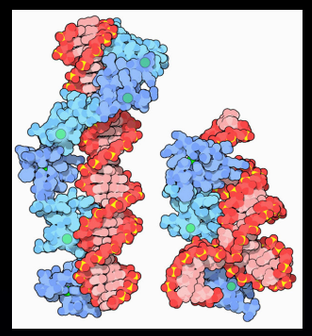


Custom report



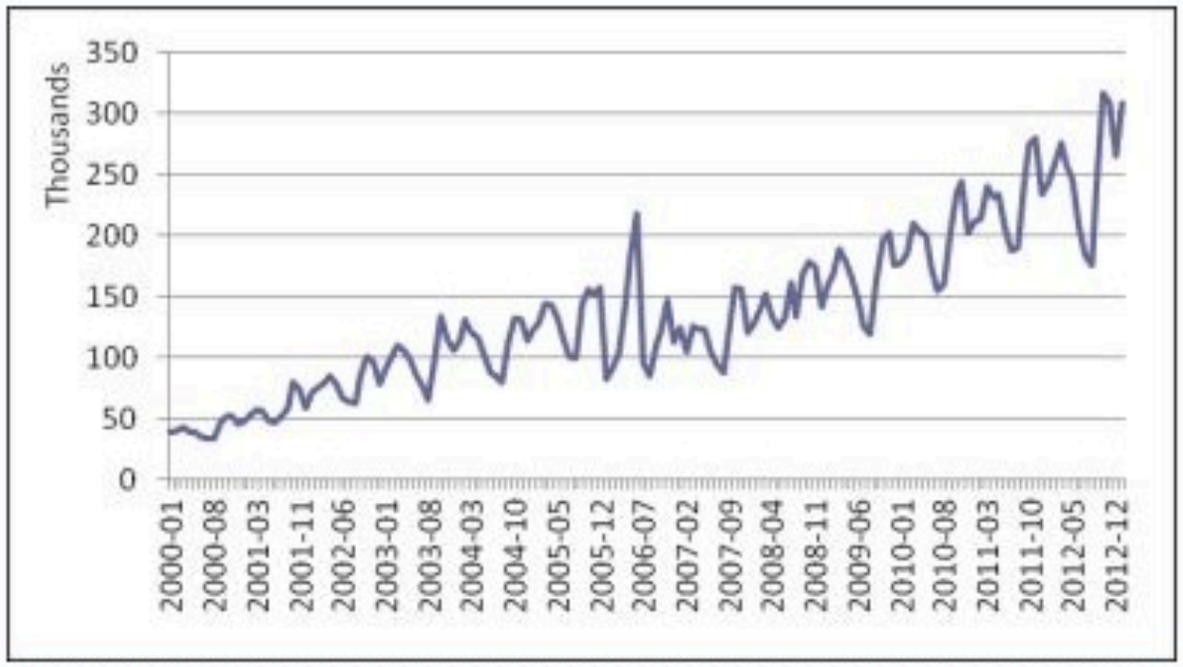
Ligand Explorer


Number of released entries
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
photosystem II
ribosome
virus
transfer RNA
myosin
nucleosome
hemoglobin
antibody
lysozyme
actin
myoglobin
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
Year

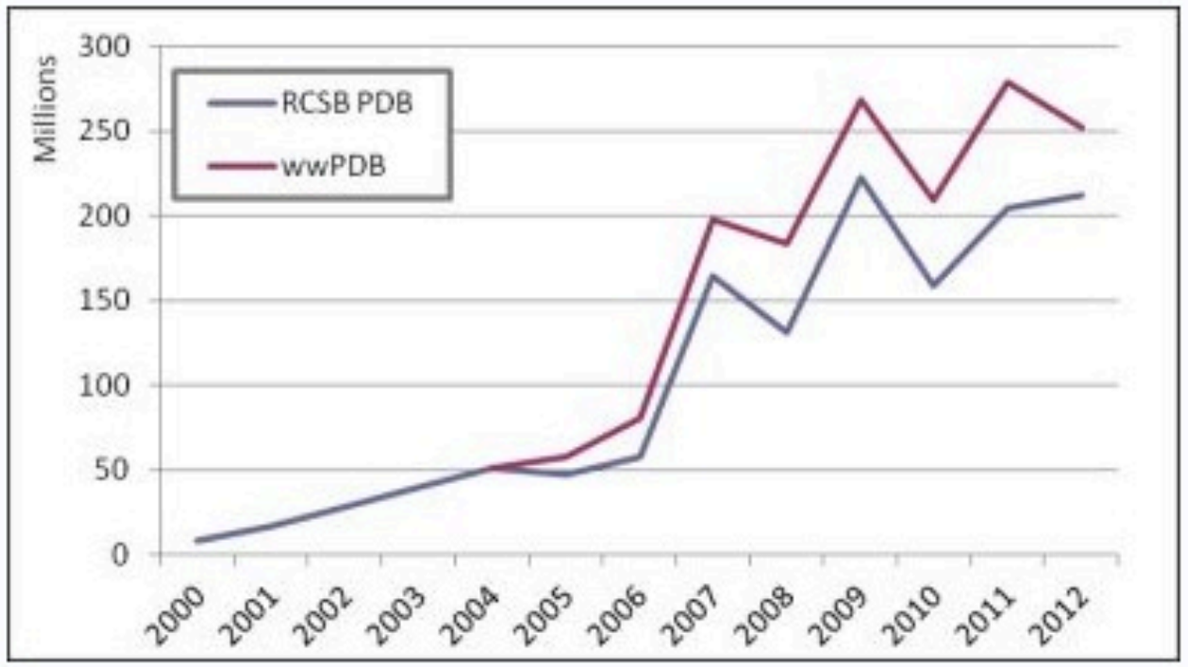


Growth in PDB Usage Over Time

RCSB PDB Website Unique Users by Month



FTP Downloads by Year



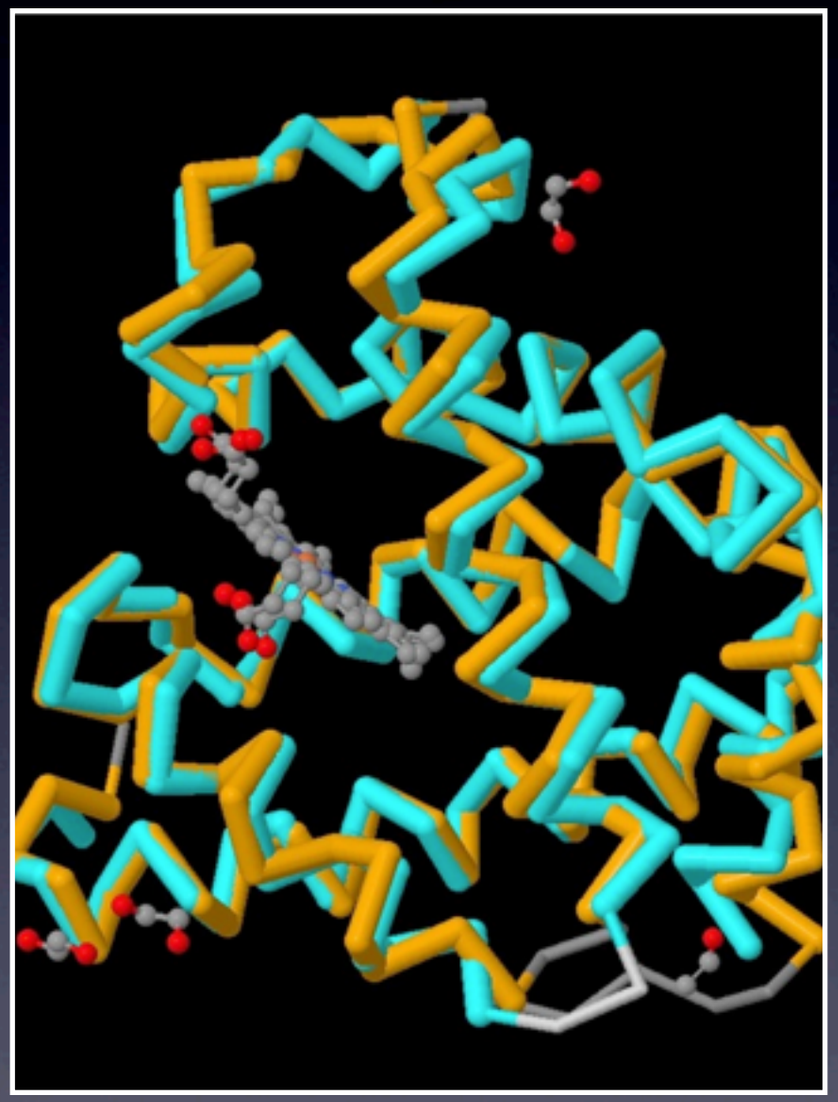
- More than 300,000 unique visitors per month
- Up to 300 concurrent users
- ~10 structures are downloaded per second 7/24/365
- Increasingly popular web services traffic

Technical Challenges

- How to provide efficient site
- Distribute data to user
- Roll out weekly updates to multiple data centers
- Availability
- How to provide interactive services
- Growing scale and complexity of structures

Our algorithms

- CPU demands
- Disk IO
- E.g. pre-calculated 3D protein structure clustering
- Parallelization, use of OSG



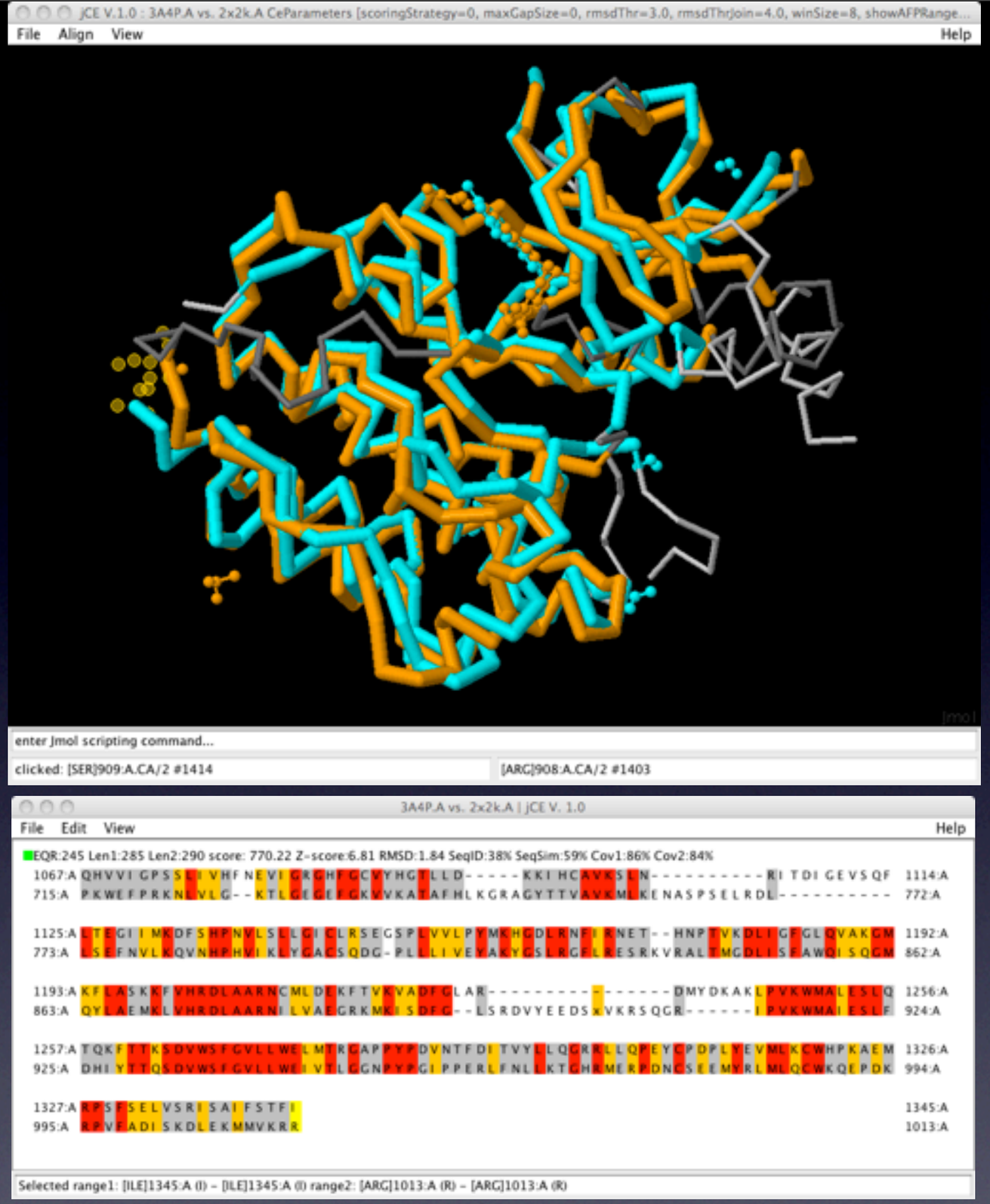
FATCAT / CE

Pairwise comparisons

Custom alignments - using Java Web Start (stand alone application)

LGPL v2 - BioJava

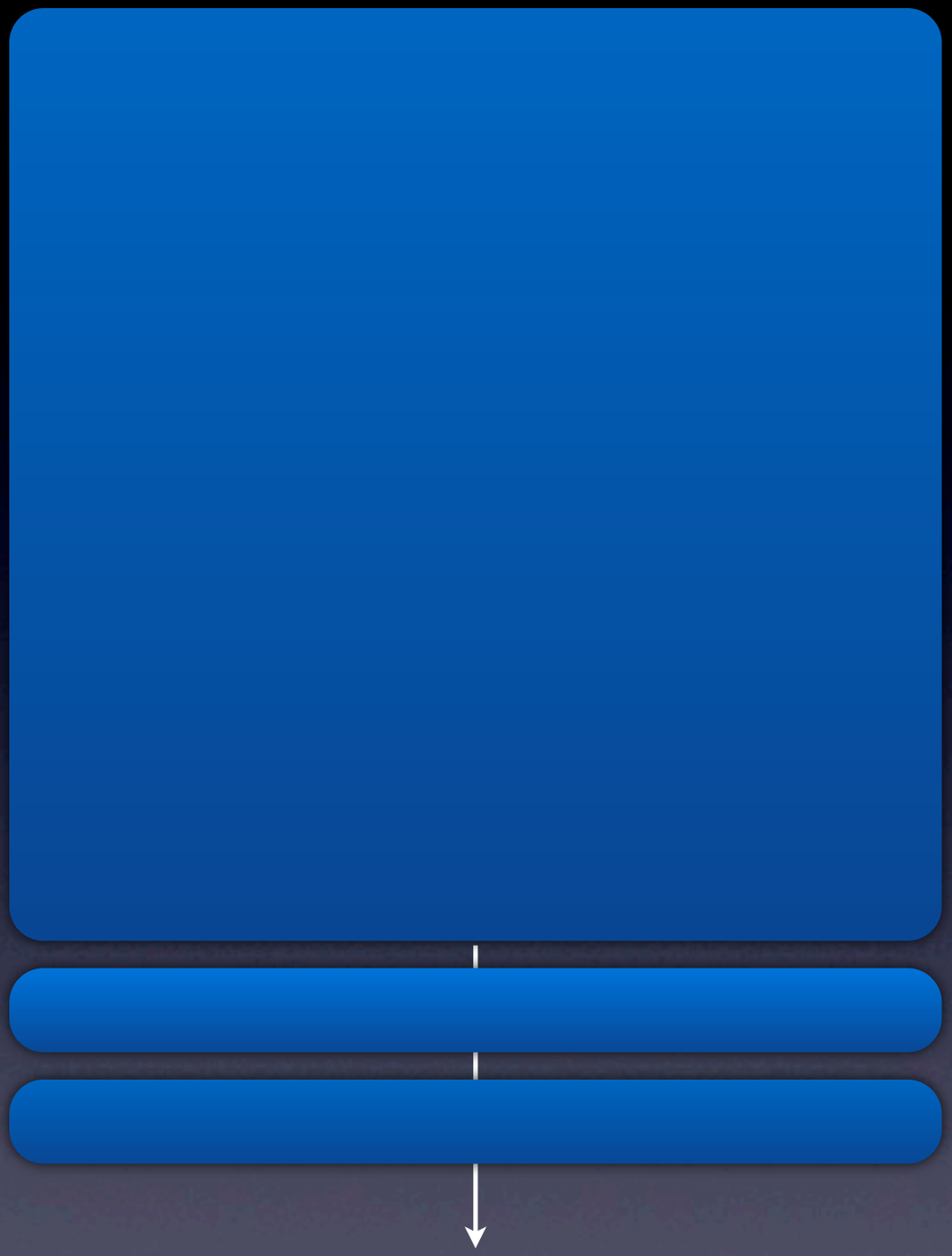
Web site - as a service



c-MET kinase domain 3A4P, RET Tyrosine Kinase 2X2K
Superimposition of ligand/inhibitor

All vs All

- Using FATCAT-rigid for All vs. All comparison...
- Well, not true all vs. all, but using representative chains...
- Step 1: Cluster sequences
- Step 2: Calculate all vs. all for ~21,000 representative chains



Initial calculation of
1 billion alignments
(~160k CPU hours)

on OSG

Incremental weekly updates
(~1 million alignments)
<1000 CPU hours

Java Clients can run anywhere

Custom Job Management

PDB

Sends out instructions to clients

Open Science Grid

Writes results to disk

Simple XML protocol

"Hello, give me something to do"

"Run these alignments"

"Here are the results"

- Efficient clients
- < 10 MB install
- Fetch data on demand from public ftp server
- Can be used by PDB users for running custom pairwise alignments and database searches (using Java Web Start)

Systematic Structural Alignment

Objective: Find novel relationships

2WUR.A (chain 1) vs. representatives of other sequence clusters (chain 2)

Rank	Result	Chain 2	Title	P-value	Score	Rmsd	Len1	Len2	%ID	%Cov1	%Cov2
1	view	2G2S.B	Green fluorescent p	0.0	478.36	0.93	226	165	96	73	100
2	view	2JAD.A	YELLOW FLUORESCI	0.0	665.32	1.01	226	346	96	100	65
3	view	3E5T.A	Red fluorescent pro	0.0	525.00	1.87	226	228	20	97	96
4	view	3EVP.A	Circular-permutate	0.0	407.39	0.35	226	223	99	61	62
5	view	3GB3.A	KillerRed	0.0	598.80	1.26	226	229	24	98	97
6	view	2G6Y.D	green fluorescent p	7.77E-16	489.59	2.22	226	214	18	93	98
7	view	3EVU.A	Myosin light chain k	2.89E-15	407.23	0.52	226	397	99	62	35
8	view	2A50.D	GFP-like non-fluores	3.06E-12	365.21	2.00	226	167	17	70	95
9	view	2G2S.A	Green fluorescent p	7.95E-10	167.91	0.22	226	64	0	27	97
10	view	1GL4.A	NIDOGEN-1	3.57E-7	295.62	3.01	226	273	9	94	78

Example: Green Fluorescent Protein

- Nidogen-1: similar 11-stranded
- beta-barrel and internal helices
- 3 Å RMSD, only 9% sequence identity
- Nidogen-1: component of basement membrane, no chromophore
- GFP and NID-1 may share common ancestor

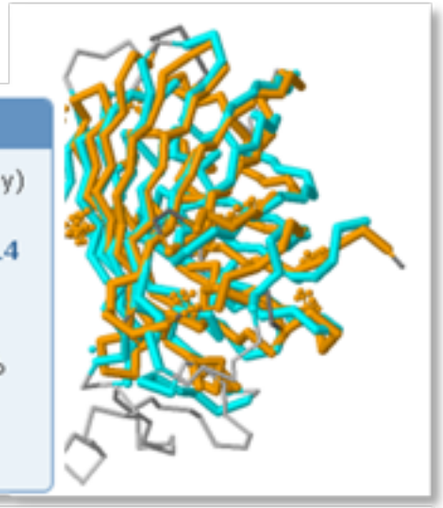
Structure Alignment Results

Alignment Details: **Query:** (colored orange/dark grey) *GREEN FLUORESCENT PROTEIN* **Subject:** (colored cyan/light grey) *NIDOGEN-1*

P-value: 3.57e-07
Score: 295.62
RMSD: 3.01
%Id: 8.8%

PDB ID: 2WUR
Chain ID: A
Length: 226
Similarity: 94%

PDB ID: 1GL4
Chain ID: A
Length: 273
Similarity: 78%



Bottlenecks

- Bottleneck no longer is CPU
- Running up to 1000 parallel jobs
- New Bottleneck: Disk IO

OSG and my environment

- Problem for many students: how to submit batch jobs
- Needed: Low entry level barrier
- Host a local UCSD - OSG tutorial for new users?
- Outreach/User support/ Helpdesk/Marketing is important
- I am on the OSG due to the Engage - outreach team
- Only later discovered the UCSD OSG group
- We have some old hardware which we would like to hook up

We are hiring

- Senior Java/Web developer

Acknowledgments

- Phil Bourne
- Peter Rose
- Chunxiao Bi
- Spencer Bliven
- Wolfgang Bluhm
- Cole Christie
- Dimitris Dimitropulous
- Alex Gramos
- Gregory Quinn
- Chris Bizon
- Adam Godzik

The RCSB PDB is managed by two members of the RCSB: Rutgers and UCSD, and is funded by NSF, NIGMS, DOE, NLM, NCI, NINDS, and NIDDK. The RCSB PDB is a member of the wwPDB.