

The screenshot shows the PDB (Protein Data Bank) website interface from 2011. At the top, there's a search bar with the text "Search articles..." and a "GO" button. Below the search bar, there are navigation links: "Home", "Browse Articles", "About", "For Readers", "For Authors and Reviewers", "Journals", "Hubs", and "PDB.org". The main content area is divided into several sections. On the left, there's a "Recent Research" section with links to articles like "Enzyme Kinetics of the Mitochondrial Deoxyribonucleoside Salvage Pathway Are Not Sufficient to Support Rapid mtDNA Replication". In the center, there's a "July 2011 Issue" section featuring a molecular structure. To the right, there's a "PDB" logo and a "PDB ID or Text" search bar. Below the search bar, there's a "Featured Molecules" section with a "Molecule of the Month: Nitric Oxide Synthase". The bottom of the page has a footer with the text "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." and "© RCSB Protein Data Bank".

# Open Data Driving Scholarly Communications in 2020

Philip E. Bourne  
UCSD

[pbourne@ucsd.edu](mailto:pbourne@ucsd.edu)

<http://www.slideshare.net/pebourne/open-data-driving-scholarly-communication-in-2020>

# My Perspective is Drawn from Being:

- A data producer
- An overseer of data curation efforts
- A database provider (PDB & IEDB)
- A data user
- Suspicious of institutional repositories
- A supporter of data publication
- Opinionated about the future

*Apologies in advance for the life sciences perspective*



This Lecture will Try and Present All  
Aspects of this Perspective

But First:

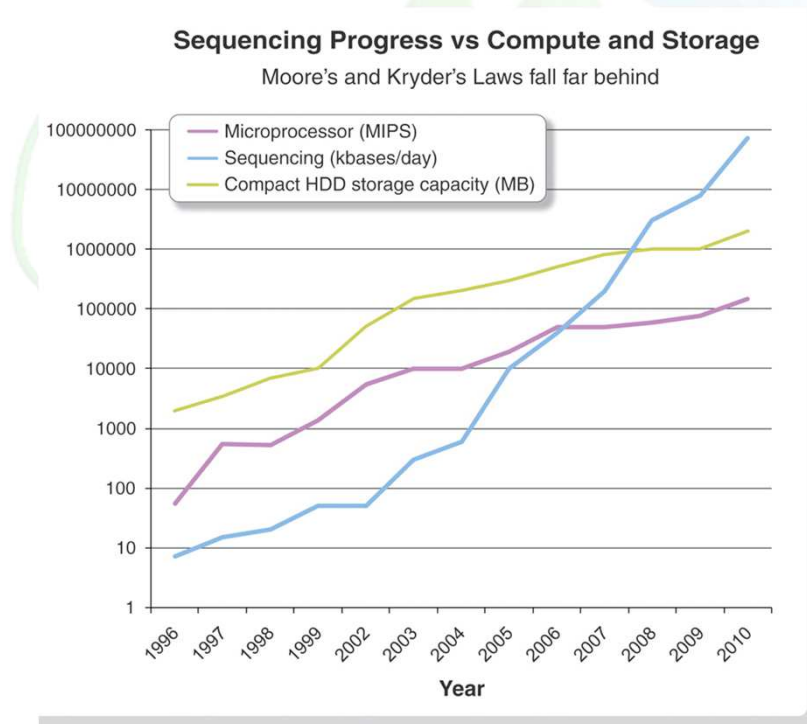
Why Open Data Are Important –

The Story of Meredith

Meredith got data the old fashioned way – she did not discover it in a broad and deep search she read the papers and bugged the authors

Imagine what she could do if data were instantly discoverable, the value quantified in some way and more simply used

# Some Thoughts as a Data Producer

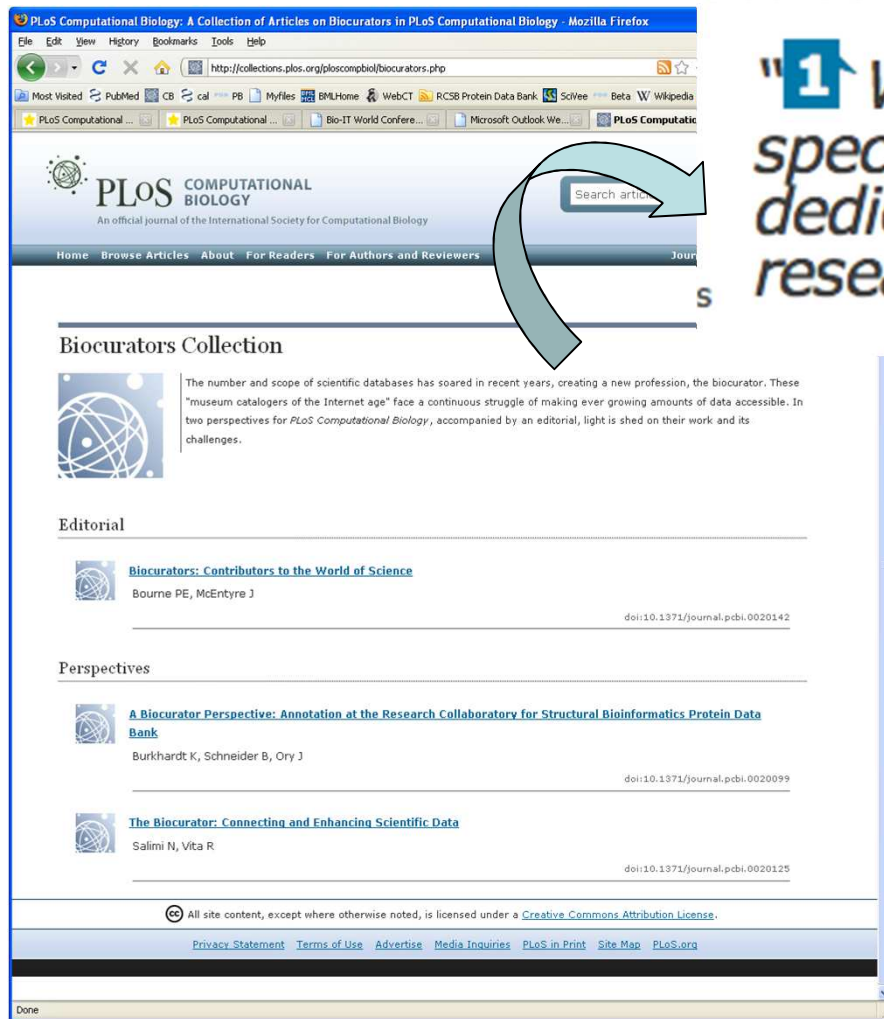


## On the Future of Genomic Data

Science 11 February 2011:  
vol. 331 no. 6018 728-729

- Its scary
- Its time to consider cost vs benefit
- Reductionism is not a dirty word
- We need to do more with the long tail

# Some Thoughts in Supporting Curation



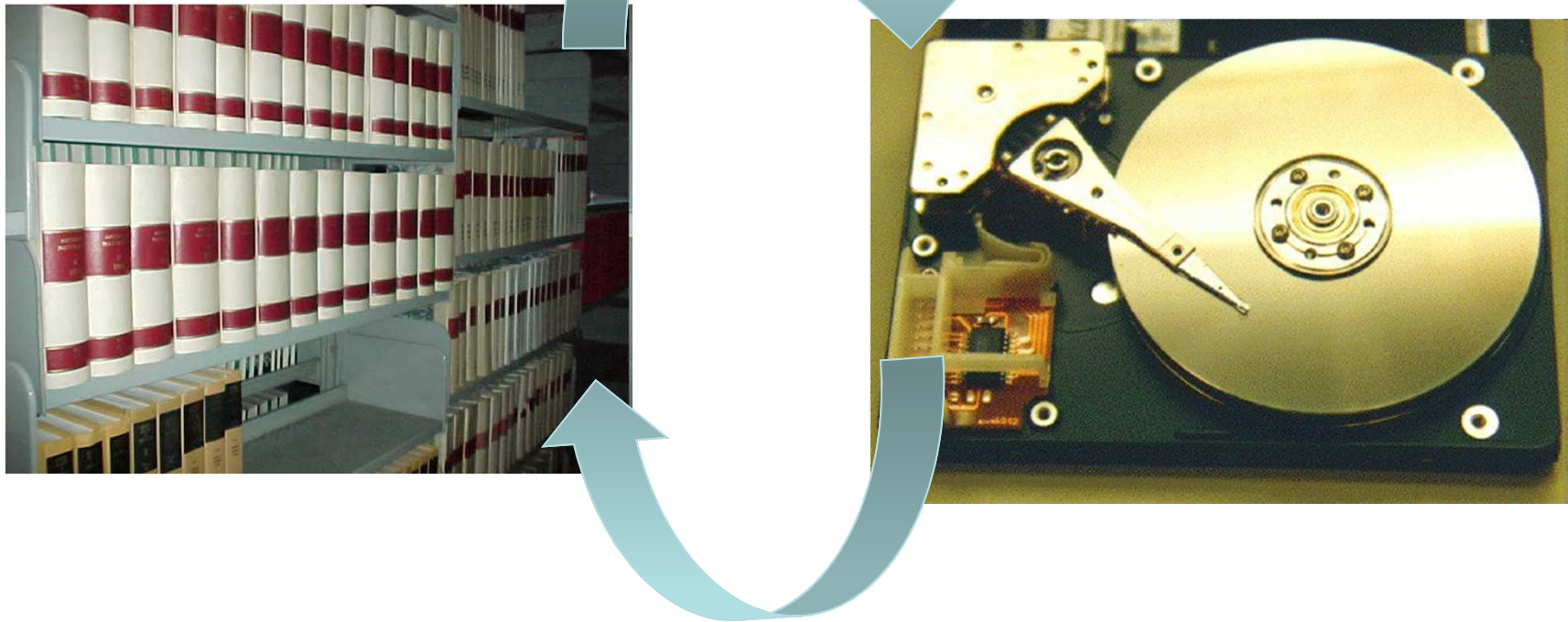
**"1** We pay homage to these special individuals who are dedicated to making our research endeavors a success."

***They really should to do more to promote themselves***

<http://collections.plos.org/ploscompbiol/biocurators.php>



# Data Curation – The Process Can be Crazy

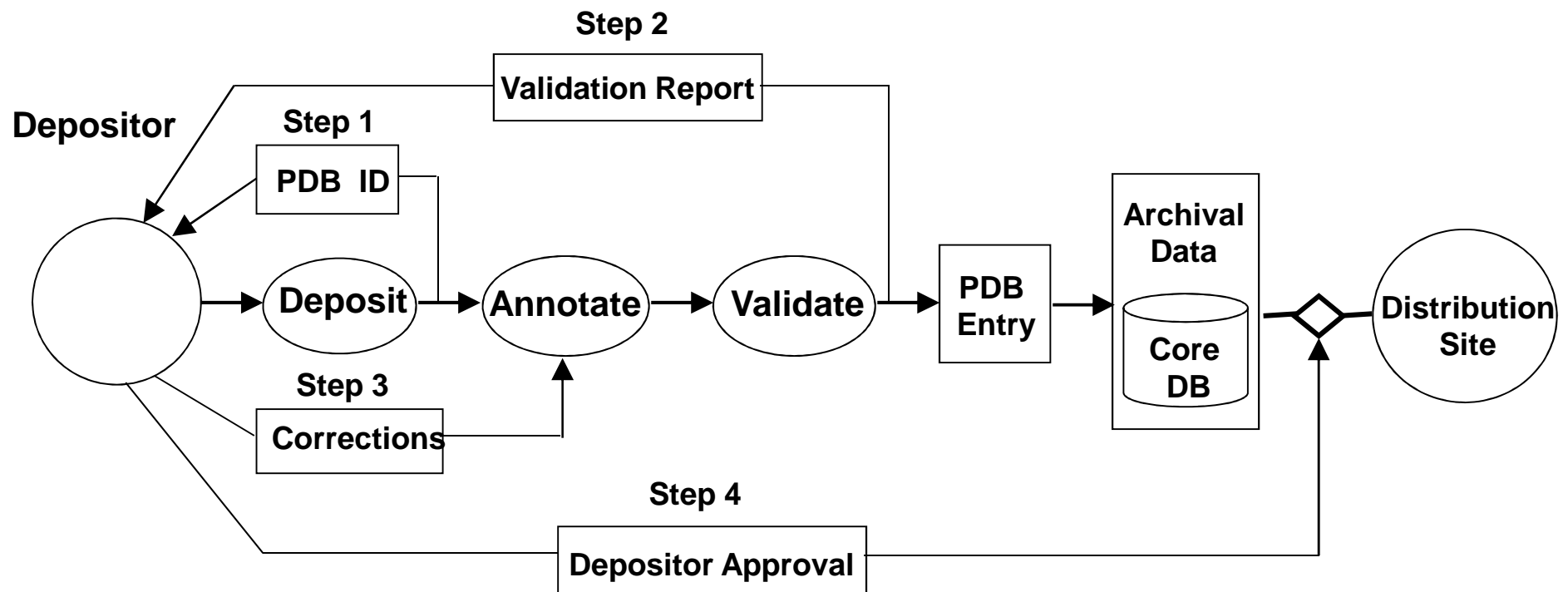


- Need new synergies between data and publication
- We will come back to this



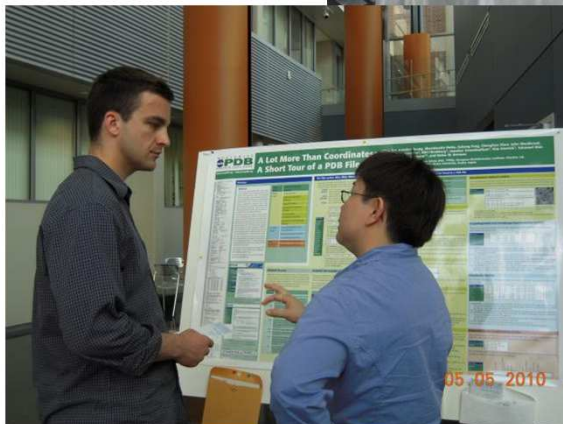
# The PDB Annotation/Validation Workflow

- Depositors do not necessarily respect the system
- Things can be too perfect



In the Future will a Biological Database Really be Different from a Biological Journal?  
PLoS Comp. Biol. 1(3) e34

# Some Happy Thoughts as a Database Provider – The PDB

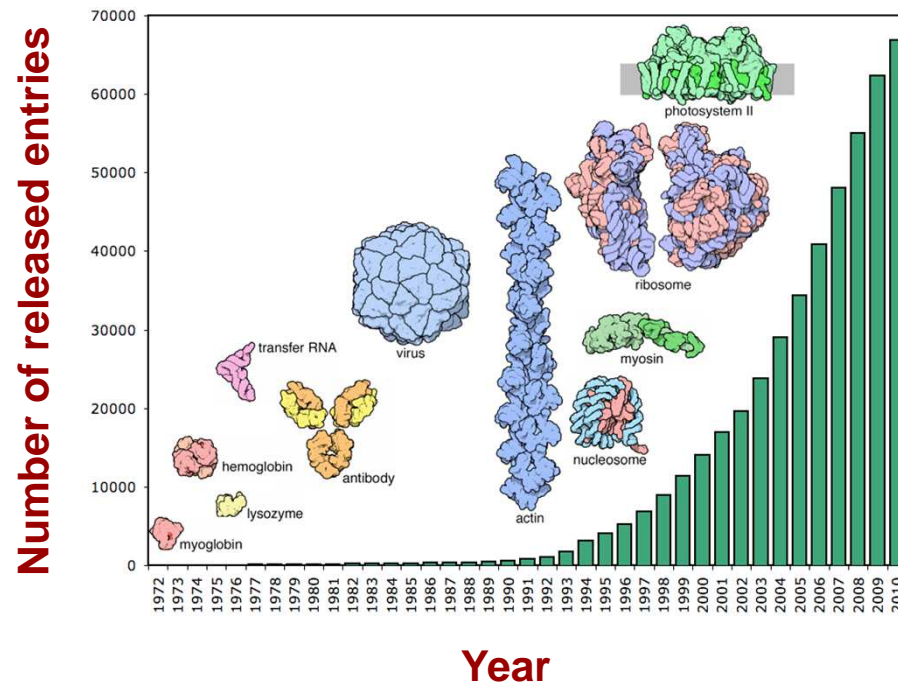


Database Provision

- Just had PDB40
- The single *community owned* worldwide repository containing structures of publically accessible biological macromolecules
- A resource distributing *worldwide* the equivalent to  $\frac{1}{4}$  the National Library of Congress each month
- A bicoastal resource
- 1TB
- Kids love it

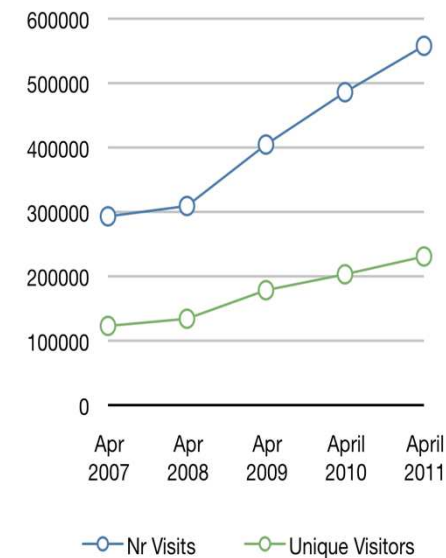
7th Int. Data Curation Conference  
Bristol UK Dec. 7, 2011

# Some Happy Thoughts as a Database Provider



We manage to handle  
Increased volume and  
complexity at a lesser cost

Website Visits and Unique Visitors



Usage increases  
and the community  
broadens

*Increasingly these define future funding,  
could it be the H-factor mistake for data?*

# Some History as a Data Provider

- About 25% of our budget has been spent on data remediation
- Support for the copy of record
- Our ontology/data model has been a critical component of our workflow and data accuracy
- Until recently the same data model was too complex to facilitate wide adoption by others that use our data

# Some History as a Data Provider

- Our data are such that we can retain redundant copies
- Data objects are discreet and we assign DOIs, but they are not used in the literature
- Constantly striving to have the user distinguish raw from derived data
- All data are not created equal but the user thinks so however hard we try

# Some Not so Happy Thoughts as a Database Provider

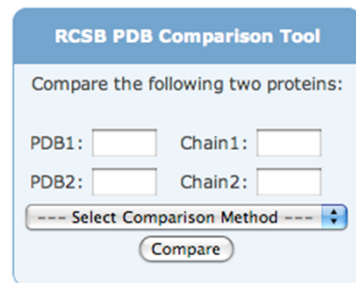
- Data are stove piped – Broad questions are difficult to answer
- Our data logs offer the means to recommend data – we do not for reasons of privacy
- Fraud may have occurred



# Trends Today as a Database Provider

- User base continues to broaden
- Constant demand for better performance (damn Google)
- Use of Web services (SOAP and now RESTful) are increasing
- The uptake on the use of widgets has been slower than I hoped

# Semantic Tagging & Widgets are a Powerful Tool to Integrate Data and Knowledge of that Data, But as Yet Not Used Much



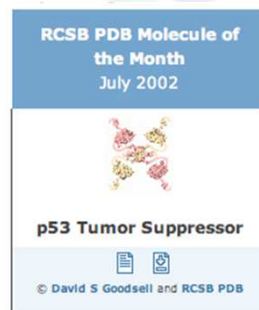
**RCSB PDB Comparison Tool**

Compare the following two proteins:

PDB1:  Chain1:

PDB2:  Chain2:

--- Select Comparison Method ---



**FOX CHASE**  
CANCER CENTER

Enter PDB code here:



PROTEIN DATA BANK

Will Widgets and Semantic Tagging Change Computational Biology?  
*PLoS Comp. Biol.* 6(2) e1000673

# Trends Today as a Database Provider

- Users are hankering after additional annotations of the data – working on database-literature integration
- Mobile use is increasing
- Web 2.0 services are in demand

# Example of Interoperability: The Database View

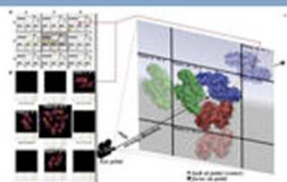
[www.rcsb.org/pdb/explore/literature.do?structureId=1TIM](http://www.rcsb.org/pdb/explore/literature.do?structureId=1TIM)

## Related Citations in PDB Entry (REMARK 1)

Show

Information provided by BioLit:

## PubMedCentral articles found to contain 1TIM ?



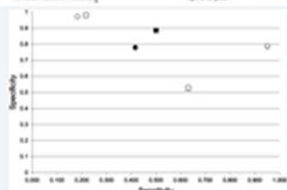
### PMG: online generation of high-quality molecular pictures and storyboarded animations

Ludovic Autin, Pierre Tufféry

*Nucleic Acids Research* 2007; 35(Web Server issue):W483-W488ff.

PubMedCentral: [1933120](#) PubMed: [17478496](#) DOI: [10.1093/nar/gkm277](#)

[Abstract](#) [Copyright](#)



### Computer-Based Screening of Functional Conformers of Proteins

Héctor Marlosti Montiel Molina, César Millán-Pacheco, Nina Pastor, Gabriel del Rio

*PLoS Computational Biology* 2008; 4(2):e1000009ff.

PubMedCentral: [2265533](#) PubMed: [18256511](#) DOI: [10.1371/journal.pcbi.1000009](#)

[Abstract](#) [Copyright](#)

## Other PDB IDs found in the above articles ?

Image	PDB ID	Structure Title	Sequence Similarity
	<b>1AON</b>	CRYSTAL STRUCTURE OF THE ASYMMETRIC CHAPERONIN COMPLEX GROEL/GROES/(ADP)7	-
	<b>1CRN</b>	WATER STRUCTURE OF A HYDROPHOBIC PROTEIN AT ATOMIC RESOLUTION. PENTAGON RINGS OF WATER MOLECULES IN CRYSTALS OF CRAMBIN	-
	<b>1A30</b>	HIV-1 PROTEASE COMPLEXED WITH A TRIPEPTIDE INHIBITOR	<i>BMC Bioinformatics</i> 2010 11:220

# Example of Interoperability – The Literature View

## App Integration in Science Direct

Result list | previous < 4 of 12 > next

PDF (339 K) | Export Citation | E-mail Article

**Article** | Figures/Tables (4)

**Journal of Molecular Biology**  
Volume 356, Issue 4, 3 March 2006, Page 1016-1024  
doi:10.1016/j.jmb.2005.11.094 | How to cite  
Copyright © 2005 Elsevier Ltd All rights reserved  
Permissions & Reprints

**Structural Insights in HIV-1 Protease NL4-3**

Holly Heaslet<sup>a</sup>, Victoria Kutilek<sup>b</sup>, Torbett<sup>b</sup> and C. David Stout<sup>a</sup>

<sup>a</sup>Department of Molecular Biology, T 92037, USA  
<sup>b</sup>Department of Molecular and Experimental Medicine, La Jolla, CA 92037, USA

**Abstract**  
The development of resistance to HIV-1 protease inhibitors is a major problem in the treatment of HIV. Therefore, it is imperative to understand the structure of HIV-1 protease NL4-3 in complex with inhibitor, TL-3. We have also obtained the crystal structures of three mutant forms of NL4-3 protease containing one (V82A), three (V82A, M46I, F53L) and six (V82A, M46I, F53L, V77I, L24I, L63P) point mutations in complex with TL-3. The three protease mutants arose sequentially under ex vivo selective pressure in the presence of TL-3, and exhibit fourfold, 11-fold, and 30-fold resistance to TL-3, respectively. This series of protease crystal structures offers insights into the biochemical and structural mechanisms by which the enzyme can overcome inhibition by TL-3 while recovering some of its native catalytic activity.

**Keywords:** HIV-1 protease; drug resistance; viral evolution; crystal structure; mutation

**Abbreviations:** HIV-1, human immunodeficiency virus type 1

**Article Outline**

**PDB Structure Viewer**

HIV-1 Protease NL4-3 in complex with inhibitor, TL-3

Release Date: 28-Feb-2006  
Exp. Method: X-RAY DIFFRACTION  
Hydrolase/hydrolase Inhibitor

Molecule: PROTEASE RETROPEPSIN  
Polymer: 1 Type: polypeptide(L)  
Chains: A  
EC#: 3.4.23.16

Molecule: TL-3 [[PHENYLMETHYLOXY-CARBONYL]-ALANINYL]-VALINYL-[PHENYL-1-HYDROXYPROP-2-YL]-AMINE  
Polymer: 2 Type: polypeptide(L)  
Chains: I  
Fragment: Half of TL-3 molecule in the asymmetric unit

**My Applications**

Add Show all apps ?

**PDB Structure Viewer**

HIV-1 Protease NL4-3 in complex with inhibitor, TL-3  
HIV-1 Protease NL4-3 1X mutant  
HIV-1 Protease NL4-3 3X mutant in complex with inhibitor, TL-3  
HIV-1 Protease NL4-3 6X mutant

About PDB Structure Viewer

**Microsoft Author Network Visualizer**

Holly Heaslet John H. Elder  
Victoria Kutilek Bruce E. Torbett  
Garrett M. Morris C. David Stout  
Ying-Chuan Lin

About Microsoft Author Network Visualizer

**Net Base Analyzer**

**Related Articles**

- Can extra-dimensional effects replace dark matter?  
*Physics Letters B*
- A generic test of modified gravity models which emulate...  
*Physics Letters B*
- Dark matter as a geometric effect in f(R) gravity  
*Astroparticle Physics*
- Planet-bound dark matter and the internal heat of Uranu...  
*Physics Letters B*
- The influence of dark matter on the motion of planets a...  
*Physics Letters A*

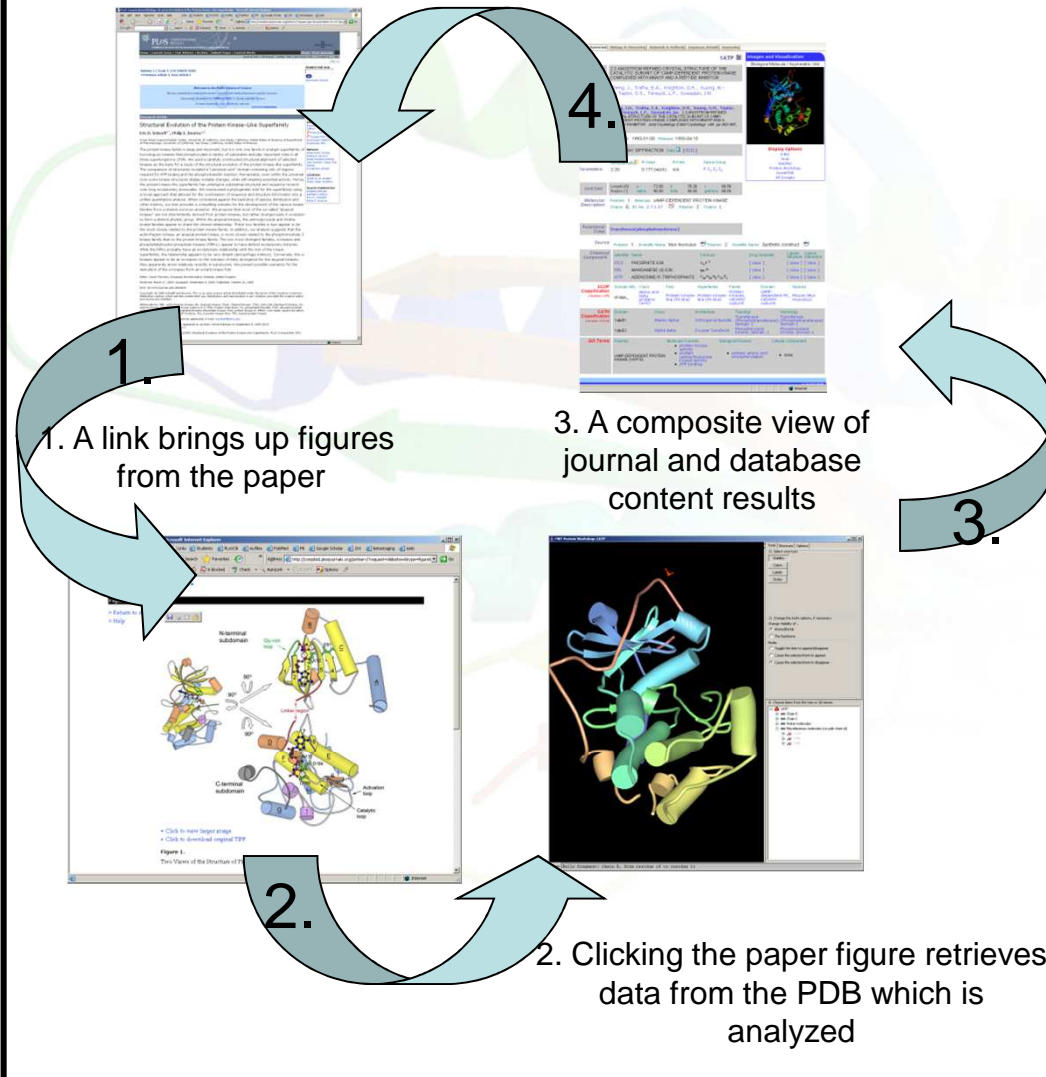
From Anita de Waard, Elsevier



# The Knowledge and Data Cycle

0. Full text of PLoS papers stored in a database

4. The composite view has links to pertinent blocks of literature text and back to the PDB



## Literature Integration – The Dream

1. User clicks on *content*
2. Metadata and webservices to data provide an *interactive view* that can be *annotated*
3. Selecting features provides a data/knowledge *mashup*
4. *Analysis* leads to new content I can *share*

PLoS Comp. Biol. 2005; 1(3): e34



# Catching our Breath...

## My Perspective is Drawn from Being:

- A data producer
  - An overseer of data curation efforts
  - A database provider (PDB & IEDB)
- A data user
  - Suspicious of institutional repositories
  - A supporter of data publication
- Opinionated about the future

# Perspective as a Data User

- Its great we are thinking more about data, but...
- Data repositories are broken
- There is a “high noon” effect
- NCBI has been a wonderful model to date...

# Data/Institutional Repositories

- *Build it and they will come* fails most of the time
- Institutional repository is an oxymoron
- NCBI works because:
  - It is an act of the US congress
  - It has strong leadership
  - It has a monopoly on the literature
  - It has IT thought out over many years

Innkeeper at the Roach Motel D. Salo 2008  
[http://muse.jhu.edu/journals/library\\_trends/v057/57.2.salo.html](http://muse.jhu.edu/journals/library_trends/v057/57.2.salo.html)

# Data/Institutional Repositories

## ■ “High Noon” Effect

- Publishers make *knowledge in* very difficult, but at least *knowledge out*, albeit limited is consistent, intuitive and easy to use
- Data repositories make *data in* and *data out* very difficult – they strive to be different when in fact users want them to be the same

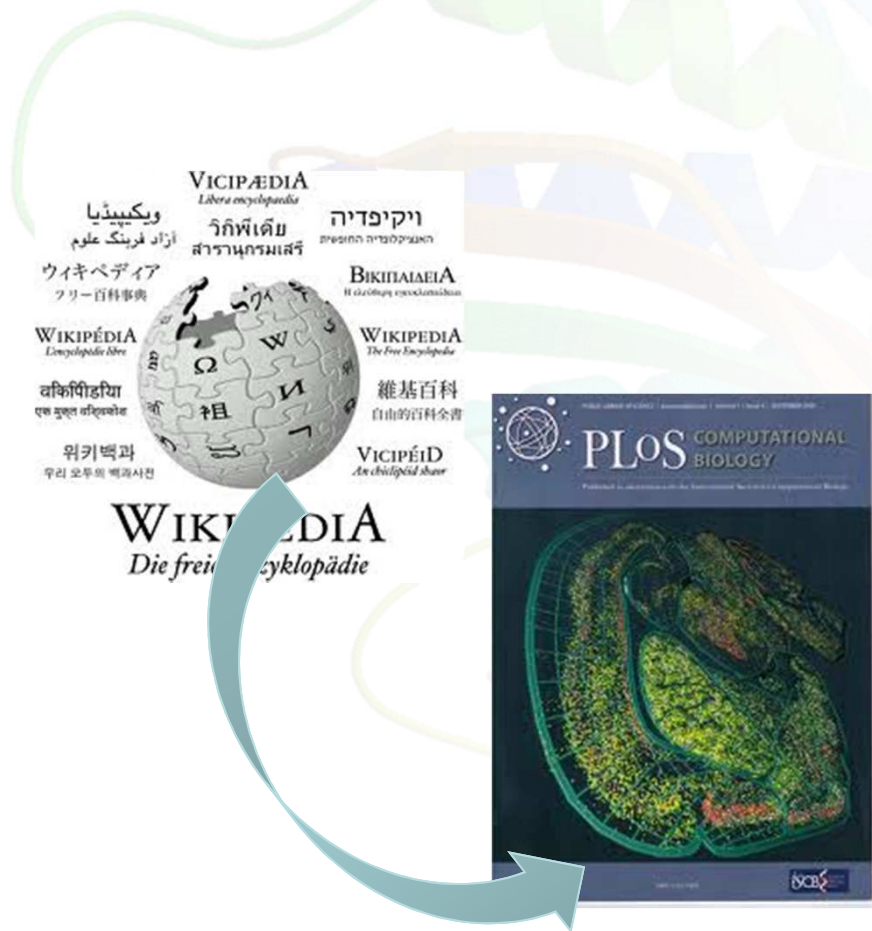
# Data and Journals

- That journals are thinking about data is good
- Dryad etc. are welcome but a stop gap measure
- Fully functional data journals will not occur without a change to the reward system
- Data papers can help shift the reward system
- Are PLoS Topic Pages a sign?

# Interim Solution:

## Use the Traditional Reward System

### The Wikipedia Experiment – Topic Pages



- Identify areas of Wikipedia that relate to the journal that are missing or stubs
- Develop a Wikipedia page in the sandbox
- Have a Topic Page Editor Review the page
- Publish the copy of record with associated rewards
- Release the living version into Wikipedia



# Catching our Breath...

## My Perspective is Drawn from Being:

- A data producer
- An overseer of data curation efforts
- A database provider (PDB & IEDB)
- A data user
- Suspicious of institutional repositories
- A supporter of data publication
- Opinionated about the future

# What Do I Want by 2020 or Earlier?

- Answer biological questions not just retrieve data
- Understand all there is to know about the availability and quality of a unit of biological data
- Operate on data in a way that is simpler, more productive, and reproducible

# What Do We Need to Do to Get There? A Data Registry?

- Individual repositories register their metadata which includes access statistics, commentary etc. – DataCite is a beginning
- Identify identical data objects and their respective metadata for comparative analysis
- Funders support registration
- Publishers support registration

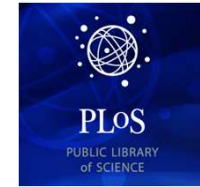
# What Do We Need to Do to Get There? An App+ Store?

- The App model
  - Think of it operating on a content base rather than a mobile device
  - Simple and consistent user interface
  - Needs to pass some quality control
  - Has a reward
- The App+ Model
  - Apps interoperate through a generic workflow interface

# Acknowledgements



**Funding Agencies:**  
NSF, NIGMS, DOE, NLM, NCI,  
NCRR, NIBIB, NINDS, NIDDK



- [www.force11.org](http://www.force11.org)
  - Tim Clark
  - Rob Dale
  - Ivan Herman
  - Ed Hovy
  - David Shotton
  - Anita de Waard
- [www.plos.org](http://www.plos.org)
- Beyond the PDF
- Many others

