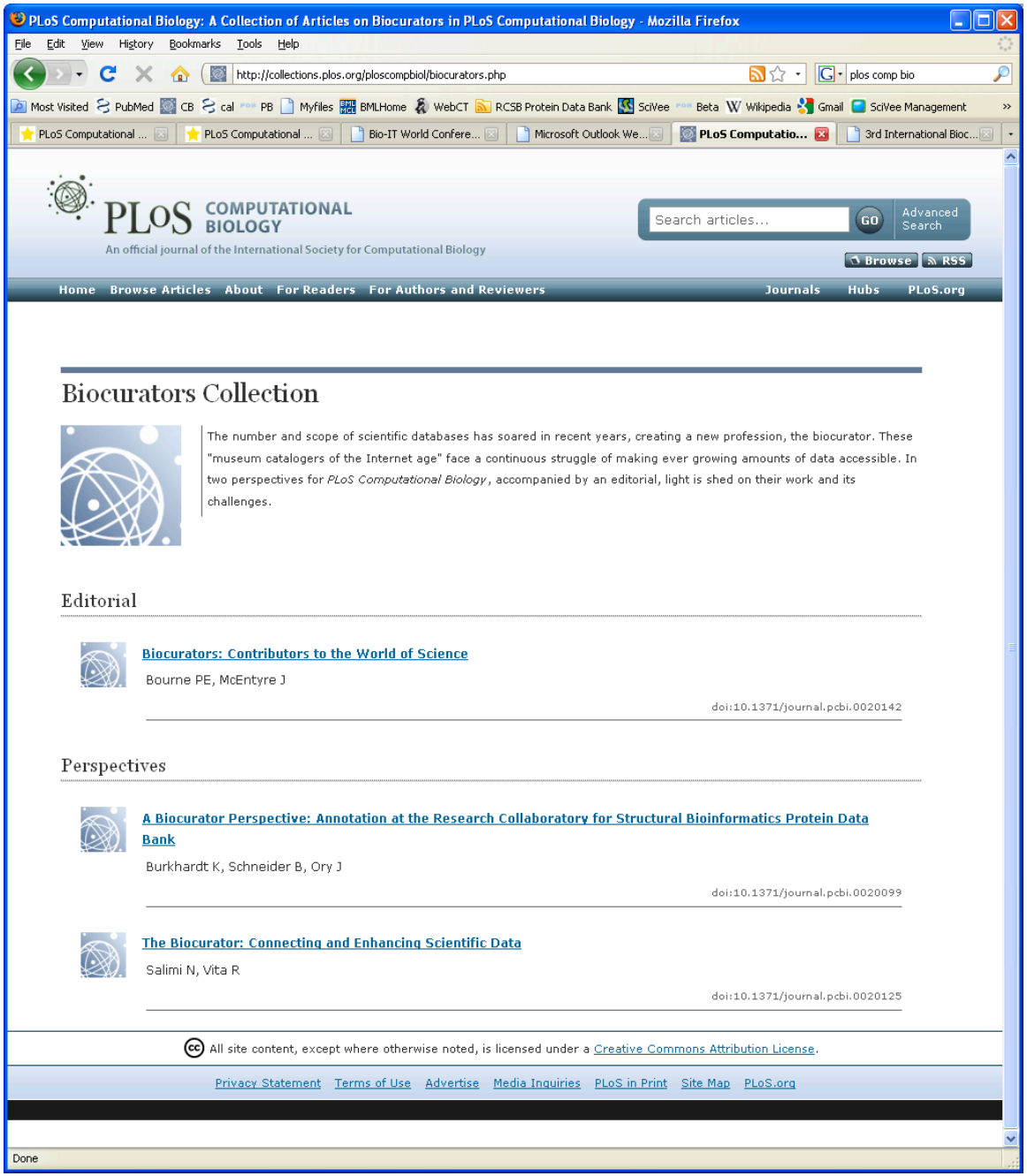
Changes in Scholarly Communication and the Potential Impact on Biocuration

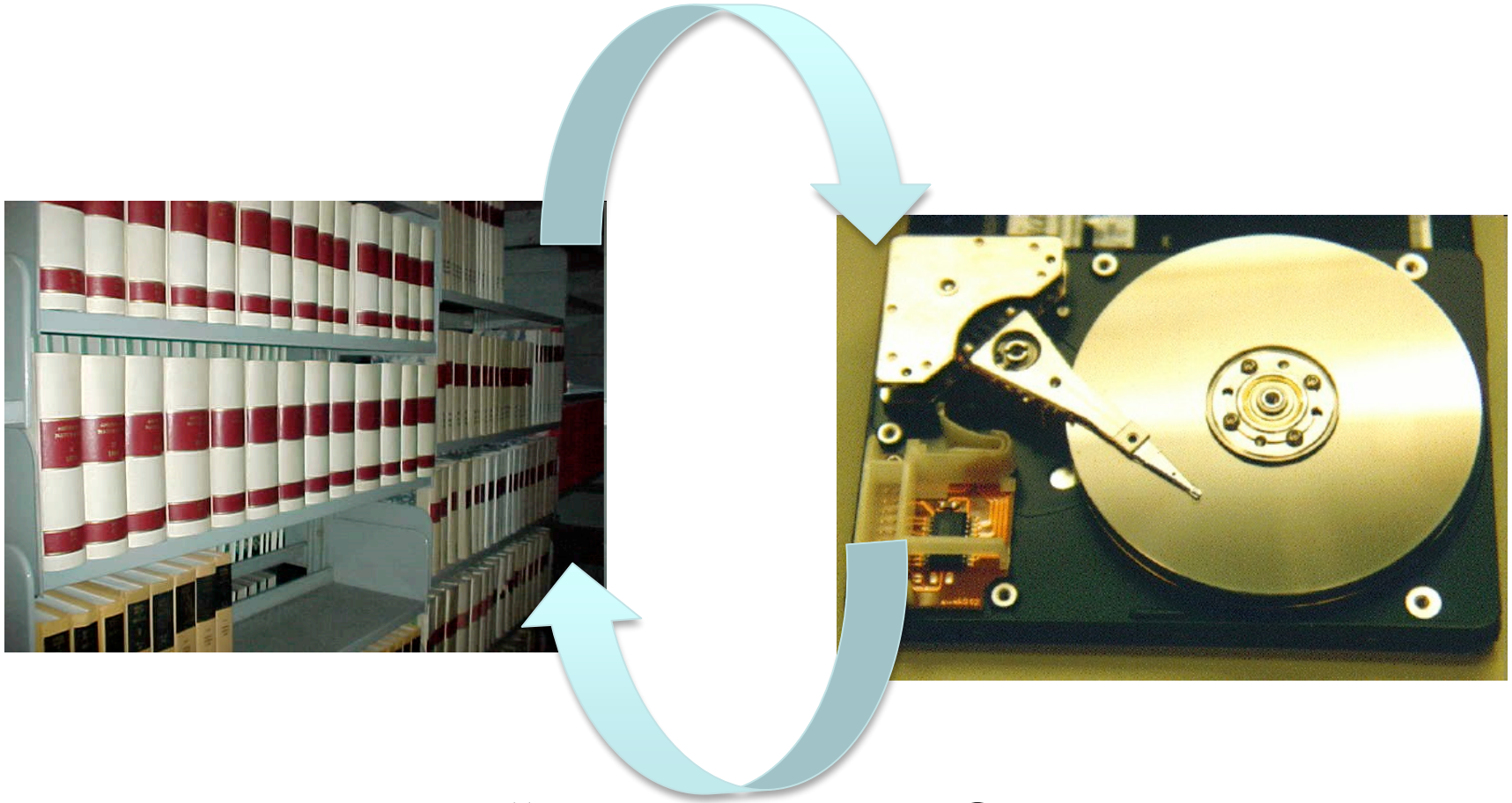
Philip E. Bourne
University of California San Diego
pbourne@ucsd.edu


Thank you!



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

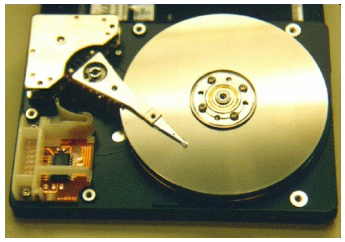
Some of What You Do – The Data Knowledge Cycle



Mostly “Value Added Curation”

The Cycle (You) Are Under Stress?

- PubMed contains 18,792,257 entries
- 50,000 papers indexed per month
- In Feb 2009:
  - 67,406,898 interactive searches were done
  - 92,216,786 entries were viewed

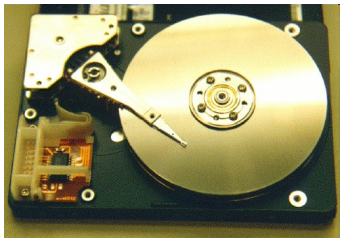


- 1078 databases reported in NAR 2008
- MetaBase http://biodatabase.org reports 2,651 entries edited 12,587 times

Data as of April 14, 2009

The Cycle Some Comparisons

- Journals have a pretty standardized interface
- Journals have a business model
- The quality is declining as numbers increase (?)
- Audience believes they are sustainable

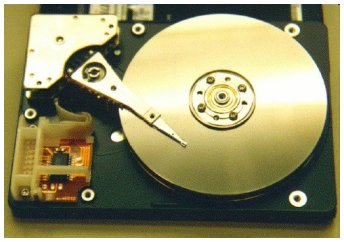


- Efforts to make the interfaces different!
- Little attempt at a business model compared to the Web 2.0 world
- Not well sustained

PLoS Comp. Biol. 2008. 4(7): e1000136

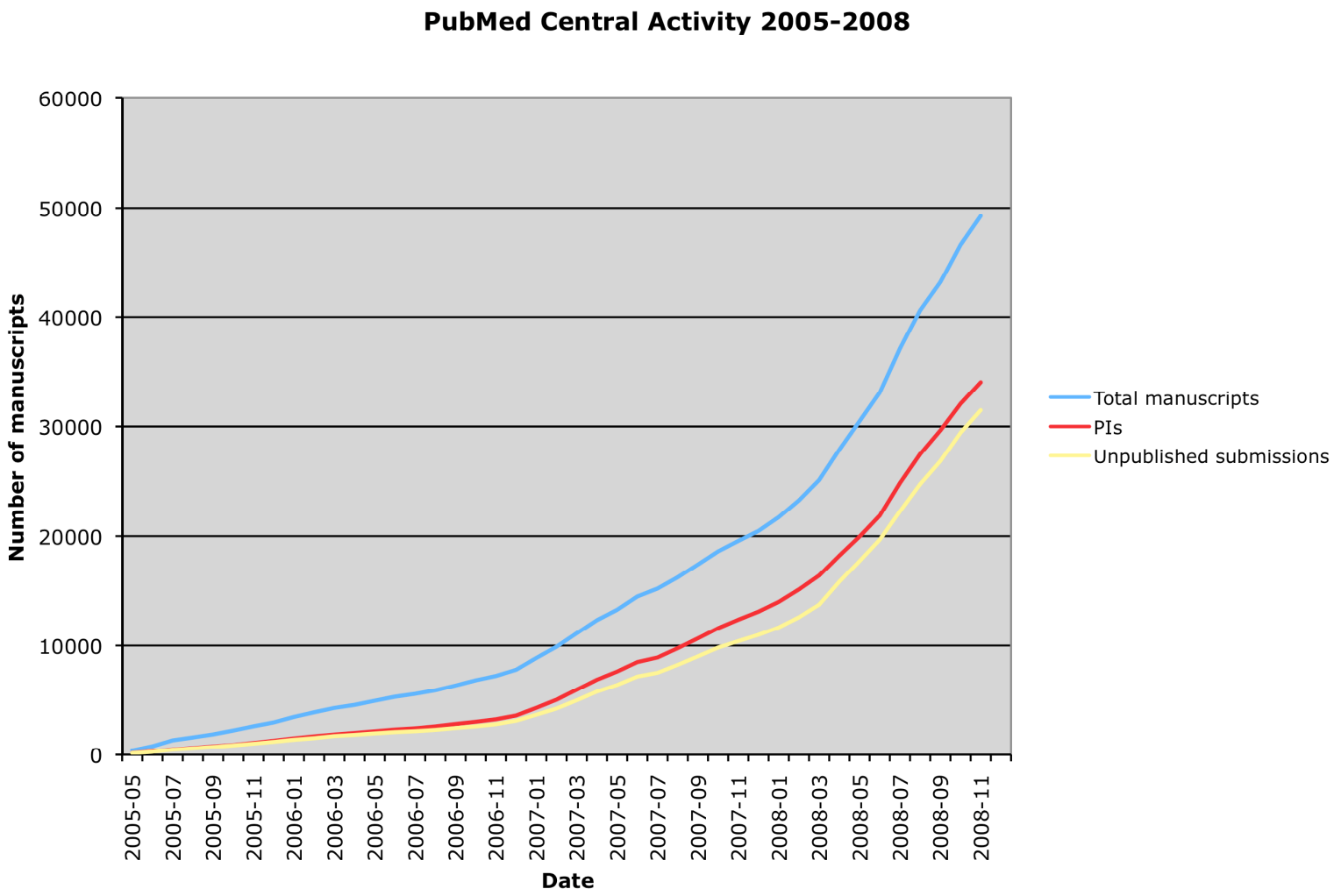
The Constituents are Changing

- New publishing models eg open access, self publishing
- Web 2.0 influence eg social networks
- Use of rich media
- The review process is failing



- Read and write eg Wikis
- New services eg restful

Growth of PubMed Central



Open Access
(Creative Commons License)

1. All published materials available on-line free to all (author pays model)
2. Unrestricted access to all published material in various formats eg XML provided attribution is given to the original author(s)
3. Copyright remains with the author

Open Access
(Creative Commons License)

1. All published materials available on-line free to all (reader pays model)
2. Unrestricted access to all published material in various formats eg XML provided attribution is given to the original author(s)
3. Copyright remains with the author

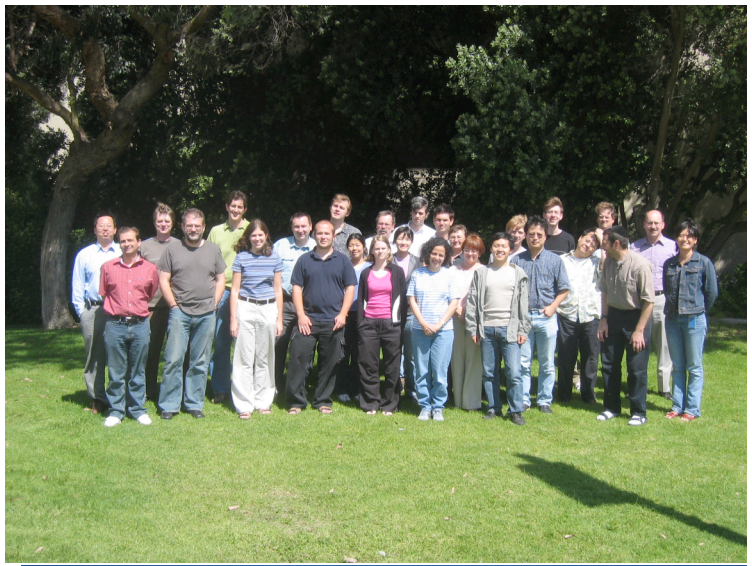
Open Access: Taking Full Advantage of the Content
PLoS Comp. Biol. 2008 4(3) e1000037

So change is afoot…
Hold that thought…

At the same time think about the notion – is a biological database really that different from a biological journal?

PLoS Comp. Biol. 2005 1(3) e34

Scholarly Communication Group



- Can we improve the way science is disseminated and comprehended?
- Through openness can we increase the number of people interested in science?

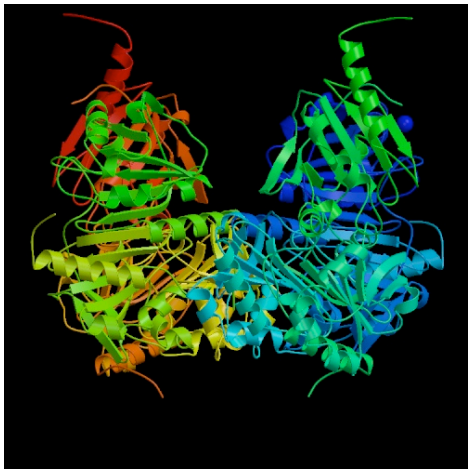
The Test Bed

http://www.plos.org/

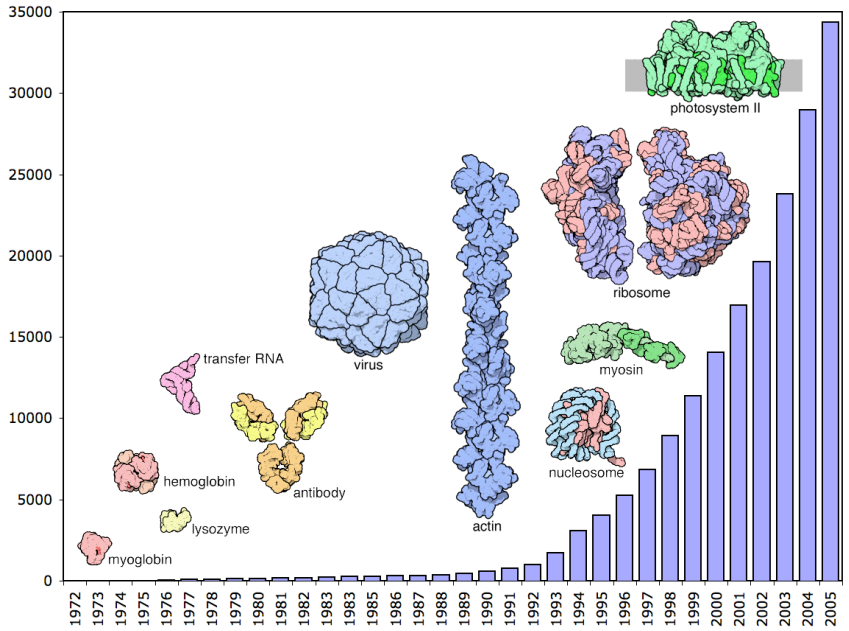
http://www.pubmedcentral.nih.gov/

WORLDWIDE PDB
PROTEIN DATA BANK

http://www.wwpdb.org/



The World Wide Protein Data Bank



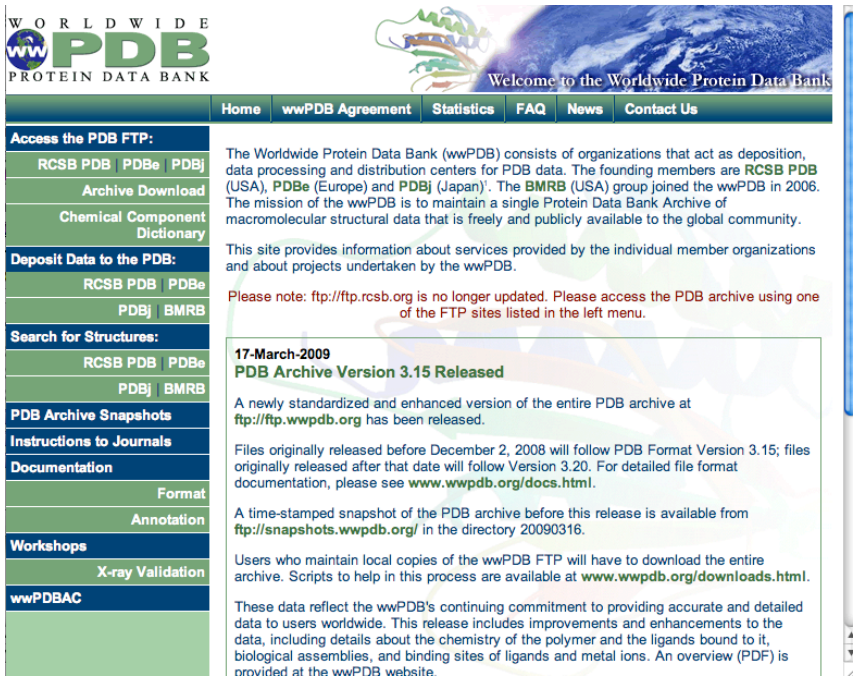
http://www.wwpdb.org

- The single worldwide repository for data on the structure of biological macromolecules
- Vital for drug discovery and the life sciences
- Over 30 years old
- Free to all

A Note in Passing

- Structural biologists have been fervent about making the data associated with their studies freely available
- For the most part they do not think the same way about the literature (knowledge) associated with the data – they hand it over without a second thought
- This latter point is true of scientists in general

The World Wide Protein Data Bank



- Paper not published unless data are deposited – strong data to literature correspondence
- Highly structured data conforming to an extensive ontology
- DOI's assigned to every structure

http://www.wwpdb.org

The PLoS Corpus

- Established in 2000
- Identified as a high quality publications (*PLoS Biology* impact factor 14.7)
- Currently 8 journals with healthy growth
- Open Access – free to all

The PLoS/PMC Corpus – Under the Hood

- Conforms well/partially to the NLM DTD – little markup of content
- PMC – some PDFs !
- The lack of conformance will come back to haunt us!

RCSB PDB
PROTEIN DATA BANK

Similar Processes Lead to Similar Resources

Author Submission via the Web

↓

Syntax Checking

↓

Review by Scientists & Editors

↓

Corrections by Author

↓

Publish – Web Accessible

Depositor Submission via the Web

↓

Syntax Checking

↓

Review by Annotators

↓

Corrections by Depositor

↓

Release – Web Accessible

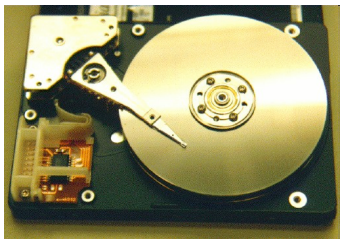
So the processes are not that dissimilar it is the final product that is perceived so differently

Even that might be changing slowly?

PLoS Comp. Biol. 2008 4(12) e1000247

We are seeing a gray area emerging between what was traditionally literature and what was traditionally databases

Examples of the Gray Area



- More data and rich media are being provided as supplemental information
- Software deposition is being encouraged

- All that you do in adding literature annotations to databases

Facilitators of Change

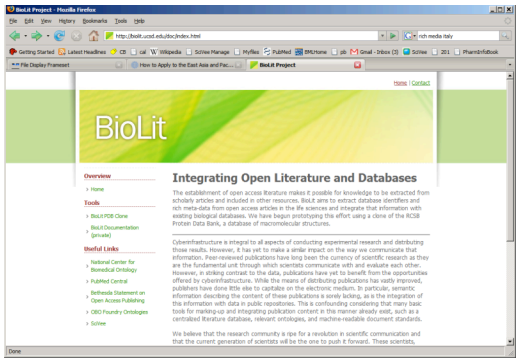
(and very useful tools for you)

Embedding semantic data during manuscript authoring

Creates an improved interface between databases and the literature

Lynn Fink

jlfink@ucsd.edu



http://biolit.ucsd.edu

Get the authors involved

- Authors are the absolute experts on the content
- More effective distribution of labor
- Add metadata before the article enters the publishing process

Word 2007 Add-in for authors

- Allows authors to add metadata as they write, before they submit the manuscript
- Authors are assisted by automated term recognition
  - OBO ontologies
  - Database IDs
- Metadata are embedded directly into the manuscript document via XML tags, OOXML format
  - Open
  - Machine-readable

- Open source, Microsoft Public License

http://www.codeplex.com/ucsdbiolit

Add-in Capabilities

- **Inline Recognition, Highlighting, and Mark-up of Informative Terms**
  - A recognized term will have a dotted, purple underline
  - Hovering generates a Smart Tag above the term
    - add mark-up for this term
    - ignore this term
    - view the term in the ontology browser
    - If a recognized term appears in more than one ontology, all instances of that term will be listed
  - Hovering over a marked-up term
    - option to apply mark-up to all recognized instances of term
    - stop recognizing a term

- **Built-in Knowledge of Ontologies and Databases**
  - Add-in provides a list of biomedical ontologies to download
  - and a list of databases for ID recognition (GenBank/RefSeq, UniProt, Protein Data Bank)
  - A user may also supply a URL to download other ontologies (soon)

- **Ontology Browser**
  - allows a user to select an ontology and then navigate through it to view terms and their relationships

Custom Metadata

- Ontologies do not contain all usages of a concept
- Add-in allows user to assign custom metadata

- Human Disease Ontology term: *Leukemia, T-Cell, HTLV-II-Associated*
- Synonym: *Atypical hairy cell leukemia (disorder)*
- Actual use in literature:
  - *hairy cell leukemia*
  - *hairy-cell leukemia*
  - *hairy T cell leukemia*
  - *T cell hairy leukemia*

Synonym mapping, disambiguation

- Inclusion of an additional set of synonyms for a term that reflect its use *in natural language*
  - Automated finding of synonyms in extant literature
  - Gather synonyms from term-mapping databases

- Incorporate a more sophisticated term recognition approach into the add-in

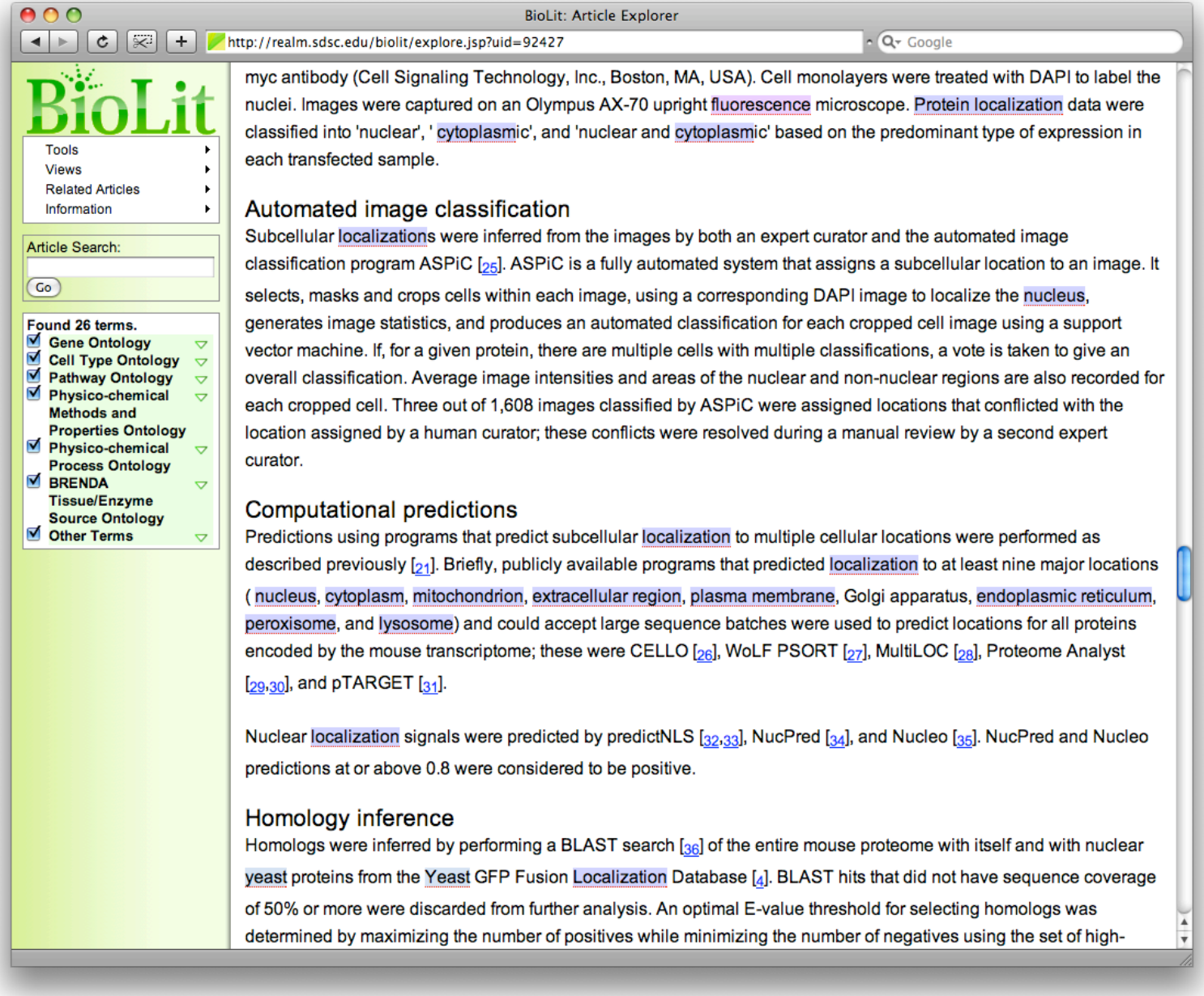
Challenges

- Author use
  - Familiarity with ontologies, terms
  - Agreement between co-authors
- End-use of semantically enriched manuscript
  - Combine with NLM XML standard
    - Article Authoring Add-in

Author Use

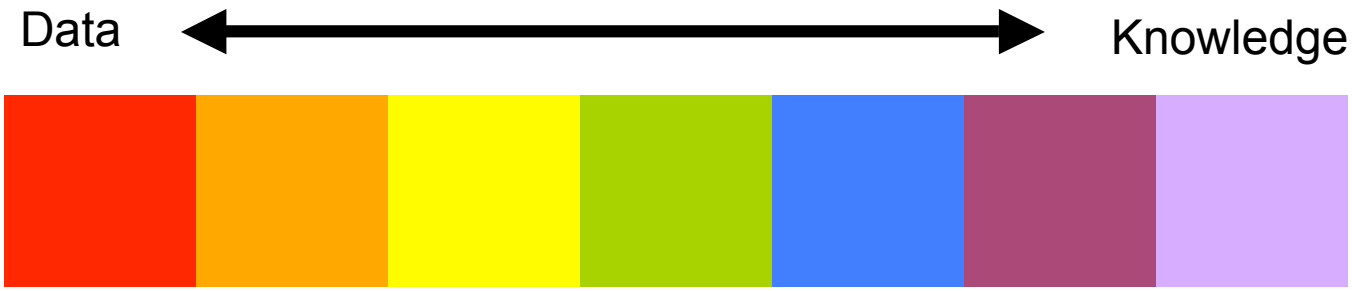
IF one or more publishers fast tracked a paper that had semantic markup I would argue it would catch on in no time

BioLit, post-processed automated



With such semantic tagging of the literature consider how it might be integrated with databases and vice-versa

First let us consider where we stand today



Database

Data Only

Knowledgebase

Data + Some Annotation

Wikis

Annotation

Datapacks

Data + Some Annotation + Some Integration

Journals

Data + Annotation

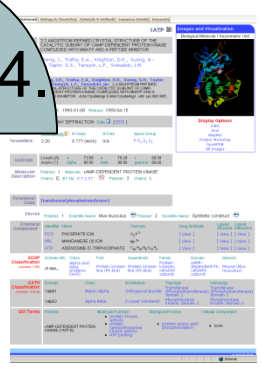
The Data – Knowledge Spectrum

BioLit: Tools for New Modes of Scientific Dissemination

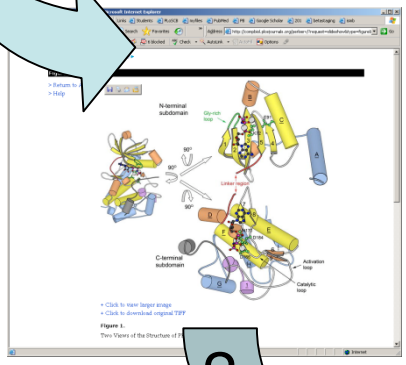
The Knowledge and Data Cycle

0. Full text of PLoS papers stored in a database

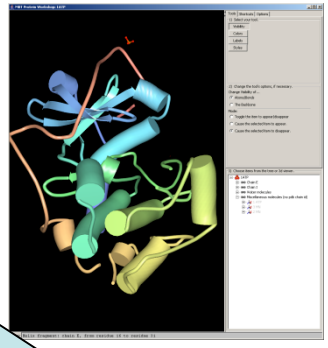
4. The composite view has links to pertinent blocks of literature text and back to the PDB



1. A link brings up figures from the paper



3. A composite view of journal and database content results



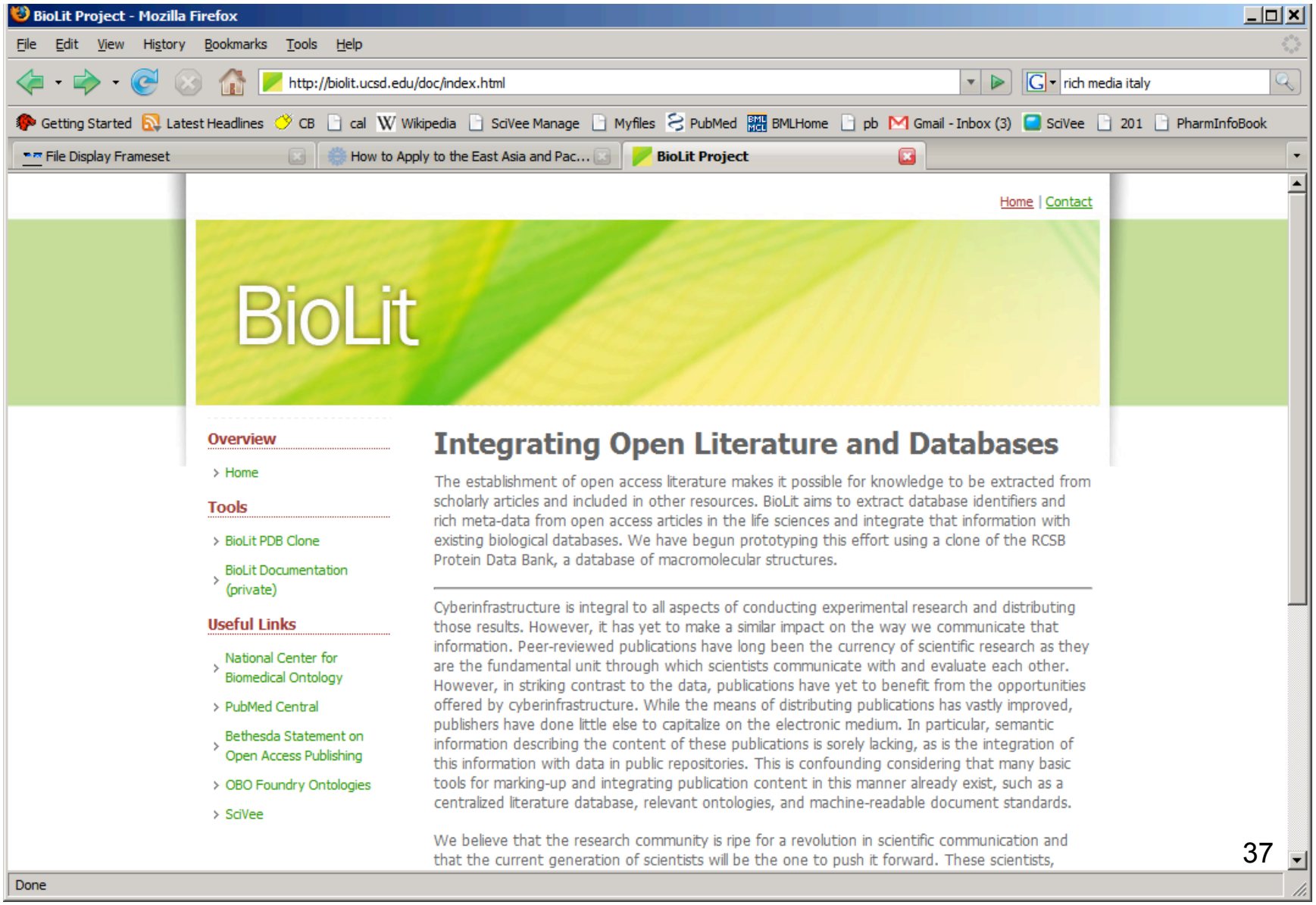
2. Clicking the paper figure retrieves data from the PDB which is analyzed

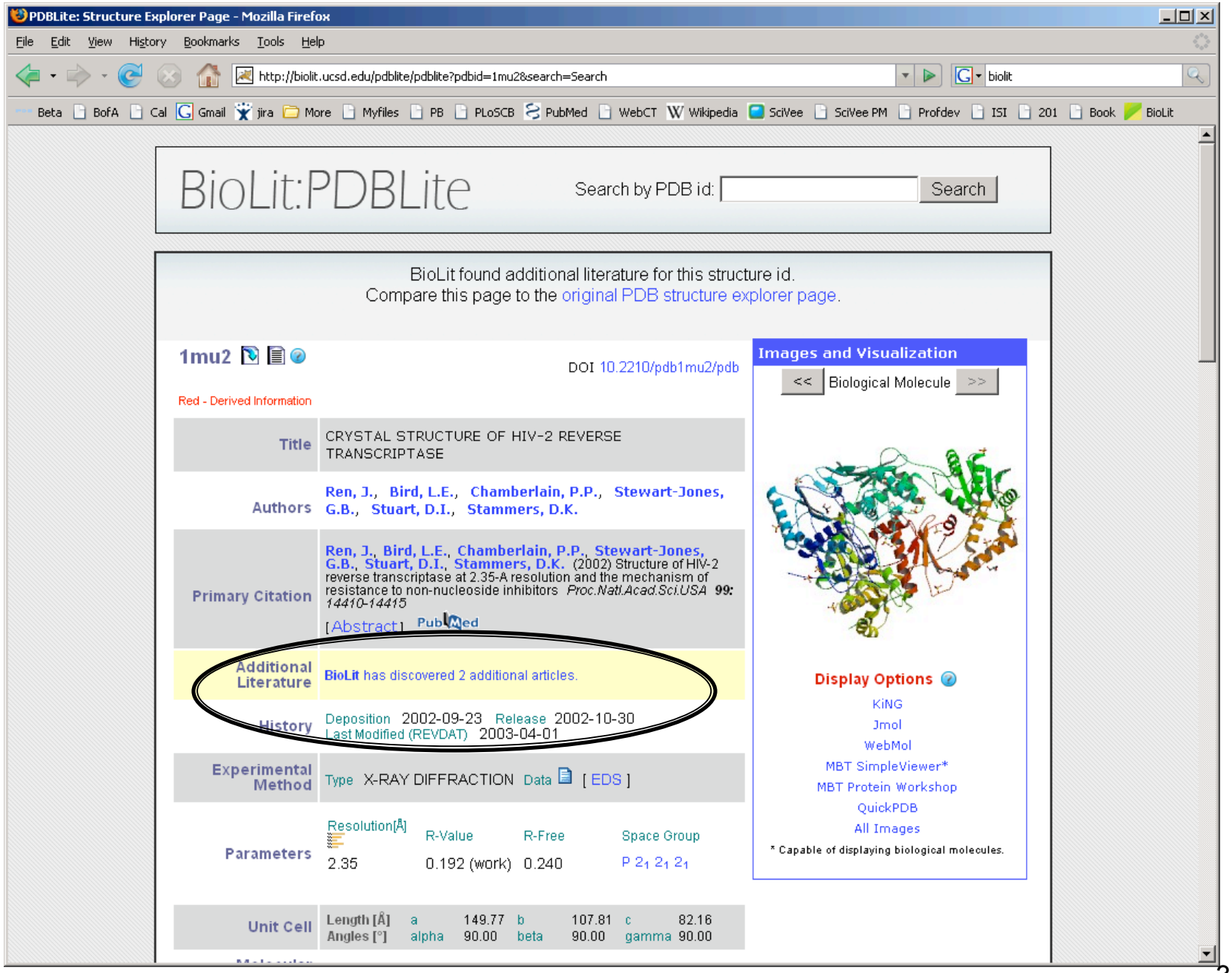
- **Biolit integrates biological literature and biological databases and includes:**
  - **A database of journal text**
  - **Authoring tools to facilitate database storage of journal text**
  - **Tools to make static tables and figures interactive**

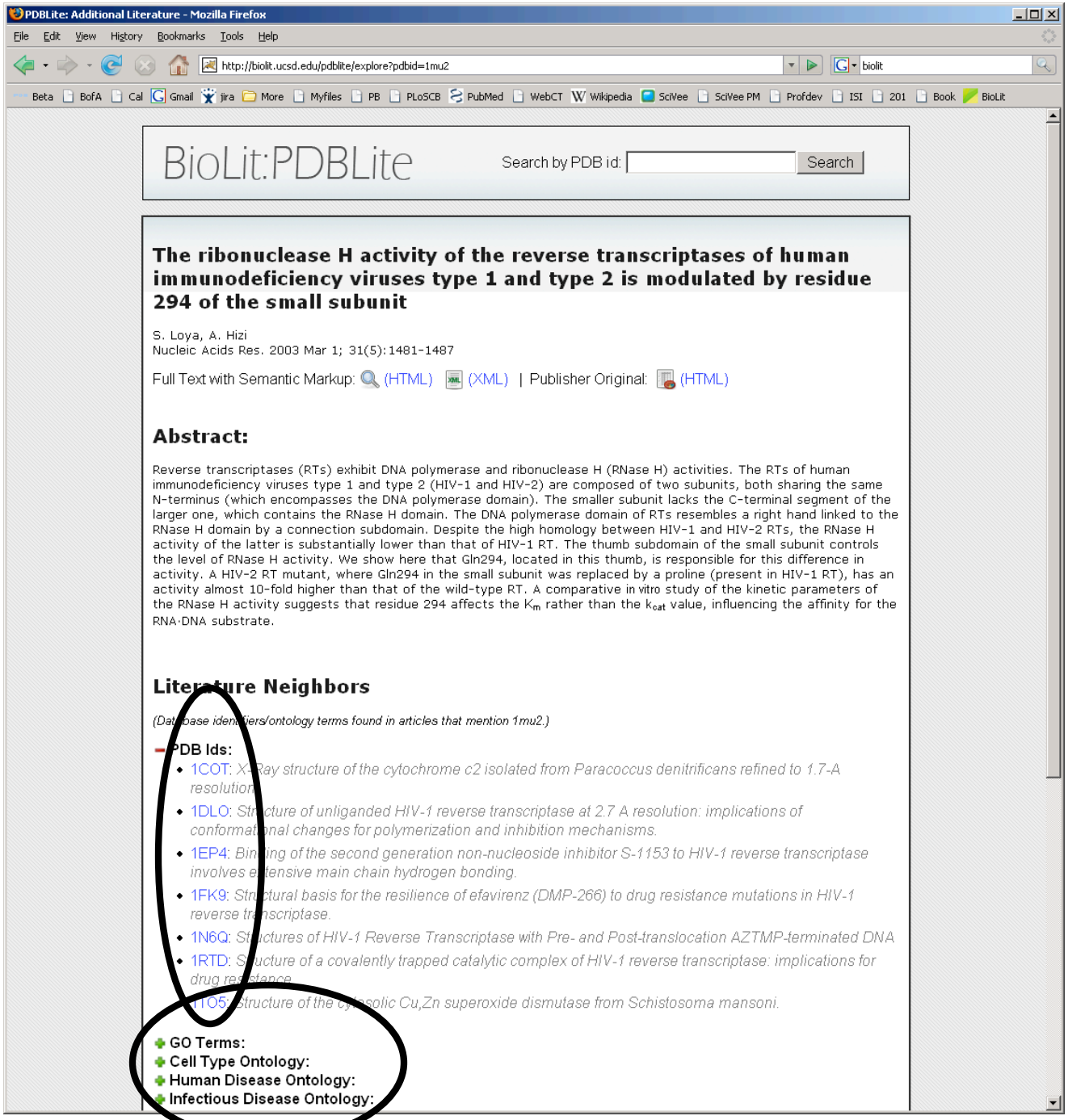
http://biolit.ucsd.edu

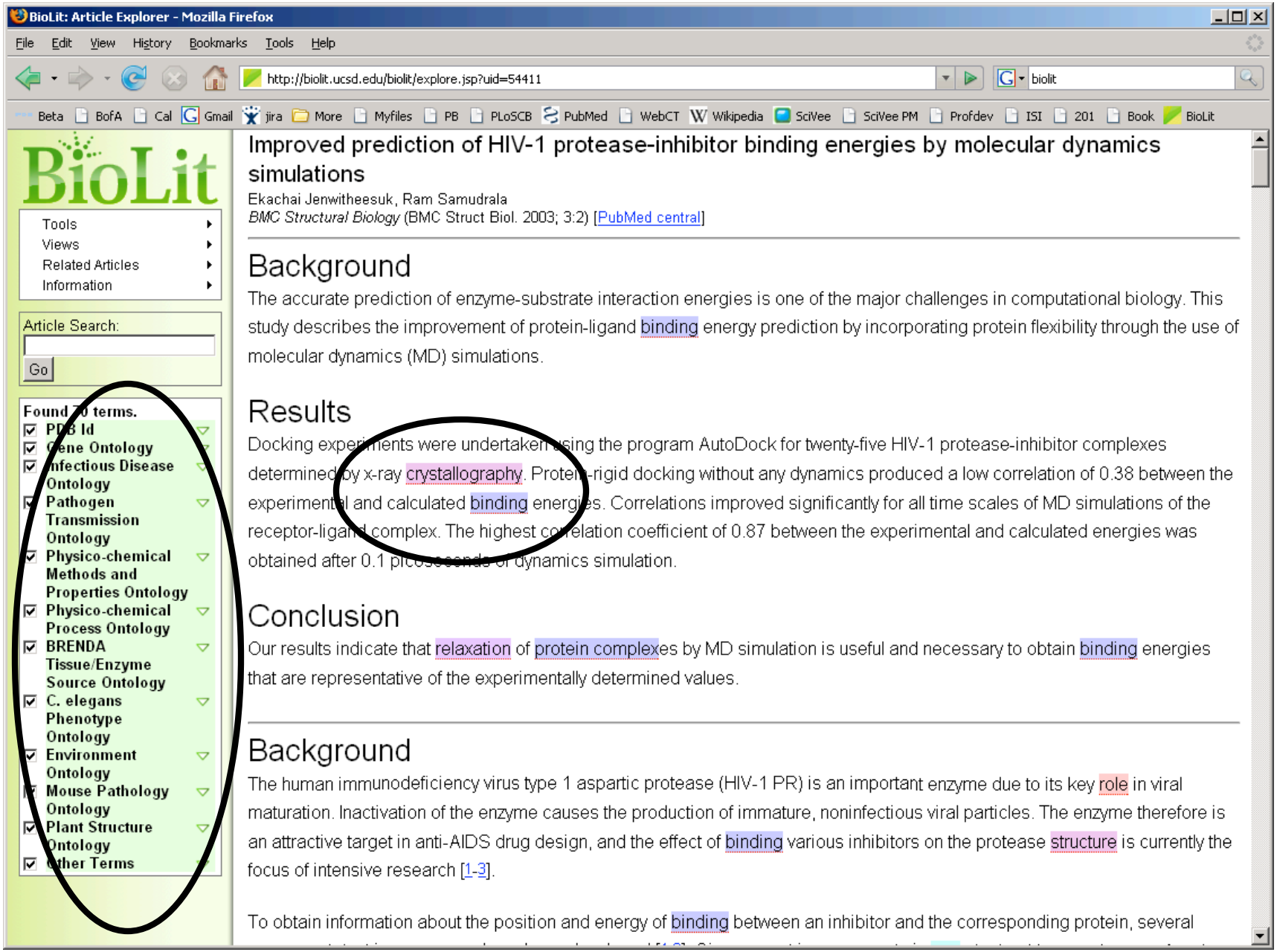
http://biolit.ucsd.edu

Nucleic Acids Research 2008 36(S2) W385-389




PDBLite: Structure Explorer Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://biolit.ucsd.edu/pdblite/pdblite?pdbid=1mu2&search=Search
biolit
Beta BofA Cal Gmail jira More Myfiles PB PLoSCB PubMed WebCT Wikipedia SciVee SciVee PM Profdev ISI 201 Book BioLit
BioLit:PDBLite
Search by PDB id:
Search
BioLit found additional literature for this structure id.
Compare this page to the original PDB structure explorer page.
1mu2
DOI 10.2210/pdb1mu2/pdb
Red - Derived Information
Title
CRYSTAL STRUCTURE OF HIV-2 REVERSE TRANSCRIPTASE
Authors
Ren, J., Bird, L.E., Chamberlain, P.P., Stewart-Jones, G.B., Stuart, D.I., Stammers, D.K.
Primary Citation
Ren, J., Bird, L.E., Chamberlain, P.P., Stewart-Jones, G.B., Stuart, D.I., Stammers, D.K. (2002) Structure of HIV-2 reverse transcriptase at 2.35-A resolution and the mechanism of resistance to non-nucleoside inhibitors Proc.Natl.Acad.Sci.USA 99: 14410-14415
[Abstract] PubMed
Additional Literature
BioLit has discovered 2 additional articles.
History
Deposition 2002-09-23 Release 2002-10-30
Last Modified (REVDAT) 2003-04-01
Experimental Method
Type X-RAY DIFFRACTION Data [EDS]
Parameters
Resolution[Å] R-Value R-Free Space Group
2.35 0.192 (work) 0.240 P 21 21 21
Unit Cell
Length [Å] a 149.77 b 107.81 c 82.16
Angles [°] alpha 90.00 beta 90.00 gamma 90.00
Images and Visualization
<< Biological Molecule >>
Display Options
KiNG
Jmol
WebMol
MBT SimpleViewer*
MBT Protein Workshop
QuickPDB
All Images
* Capable of displaying biological molecules.


PDBLite: Additional Literature - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://biolit.ucsd.edu/pdblite/explore?pdbid=1mu2
BioLit:PDBLite
Search by PDB id:
Search
The ribonuclease H activity of the reverse transcriptases of human immunodeficiency viruses type 1 and type 2 is modulated by residue 294 of the small subunit
S. Loya, A. Hizi
Nucleic Acids Res. 2003 Mar 1; 31(5):1481-1487
Full Text with Semantic Markup: (HTML) (XML) | Publisher Original: (HTML)
Abstract:
Reverse transcriptases (RTs) exhibit DNA polymerase and ribonuclease H (RNase H) activities. The RTs of human immunodeficiency viruses type 1 and type 2 (HIV-1 and HIV-2) are composed of two subunits, both sharing the same N-terminus (which encompasses the DNA polymerase domain). The smaller subunit lacks the C-terminal segment of the larger one, which contains the RNase H domain. The DNA polymerase domain of RTs resembles a right hand linked to the RNase H domain by a connection subdomain. Despite the high homology between HIV-1 and HIV-2 RTs, the RNase H activity of the latter is substantially lower than that of HIV-1 RT. The thumb subdomain of the small subunit controls the level of RNase H activity. We show here that Gln294, located in this thumb, is responsible for this difference in activity. A HIV-2 RT mutant, where Gln294 in the small subunit was replaced by a proline (present in HIV-1 RT), has an activity almost 10-fold higher than that of the wild-type RT. A comparative in vitro study of the kinetic parameters of the RNase H activity suggests that residue 294 affects the K_m rather than the k_{cat} value, influencing the affinity for the RNA·DNA substrate.
Literature Neighbors
(Database identifiers/ontology terms found in articles that mention 1mu2.)
PDB Ids:
1COT: X-Ray structure of the cytochrome c2 isolated from Paracoccus denitrificans refined to 1.7-A resolution
1DLO: Structure of unliganded HIV-1 reverse transcriptase at 2.7 A resolution: implications of conformational changes for polymerization and inhibition mechanisms.
1EP4: Binding of the second generation non-nucleoside inhibitor S-1153 to HIV-1 reverse transcriptase involves extensive main chain hydrogen bonding.
1FK9: Structural basis for the resilience of efavirenz (DMP-266) to drug resistance mutations in HIV-1 reverse transcriptase.
1N6Q: Structures of HIV-1 Reverse Transcriptase with Pre- and Post-translocation AZTMP-terminated DNA
1RTD: Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance
1TO5: Structure of the cytosolic Cu,Zn superoxide dismutase from Schistosoma mansoni.
GO Terms:
Cell Type Ontology:
Human Disease Ontology:
Infectious Disease Ontology:


BioLit: Article Explorer - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://biolit.ucsd.edu/biolit/explore.jsp?uid=54411
biolit
Beta BofA Cal Gmail jira More Myfiles PB PLoSCB PubMed WebCT Wikipedia SciVee SciVee PM Profdev ISI 201 Book BioLit
BioLit
Tools
Views
Related Articles
Information
Article Search:
Go
Found 20 terms.
PDB Id
Gene Ontology
Infectious Disease Ontology
Pathogen Transmission Ontology
Physico-chemical Methods and Properties Ontology
Physico-chemical Process Ontology
BRENDA Tissue/Enzyme Source Ontology
C. elegans Phenotype Ontology
Environment Ontology
Mouse Pathology Ontology
Plant Structure Ontology
Other Terms
Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations
Ekachai Jenwitheesuk, Ram Samudrala
BMC Structural Biology (BMC Struct Biol. 2003; 3:2) [PubMed central]
Background
The accurate prediction of enzyme-substrate interaction energies is one of the major challenges in computational biology. This study describes the improvement of protein-ligand binding energy prediction by incorporating protein flexibility through the use of molecular dynamics (MD) simulations.
Results
Docking experiments were undertaken using the program AutoDock for twenty-five HIV-1 protease-inhibitor complexes determined by x-ray crystallography. Protein-rigid docking without any dynamics produced a low correlation of 0.38 between the experimental and calculated binding energies. Correlations improved significantly for all time scales of MD simulations of the receptor-ligand complex. The highest correlation coefficient of 0.87 between the experimental and calculated energies was obtained after 0.1 picoseconds of dynamics simulation.
Conclusion
Our results indicate that relaxation of protein complexes by MD simulation is useful and necessary to obtain binding energies that are representative of the experimentally determined values.
Background
The human immunodeficiency virus type 1 aspartic protease (HIV-1 PR) is an important enzyme due to its key role in viral maturation. Inactivation of the enzyme causes the production of immature, noninfectious viral particles. The enzyme therefore is an attractive target in anti-AIDS drug design, and the effect of binding various inhibitors on the protease structure is currently the focus of intensive research [1-3].
To obtain information about the position and energy of binding between an inhibitor and the corresponding protein, several

BioLit: Article Explorer - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://biolit.ucsd.edu/biolit/explore.jsp?uid=54411

Beta BofA Cal Gmail jira More Myfiles PB PLoSCB PubMed WebCT Wikipedia SciVee SciVee PM Profdev ISI 201 Book BioLit

BioLit

Tools
Views
Related Articles
Information

Article Search:

Go

Found 70 terms.

- PDB Id
  - 1HVI (1)
  - 1HVJ (1)
  - 1HVK (1)
  - 1HVL (1)
  - 4PHV (2)
- Gene Ontology
- Infectious Disease Ontology
- Pathogen Transmission Ontology
- Physico-chemical Methods and Properties Ontology
- Physico-chemical Process Ontology
- BRENDA Tissue/Enzyme Source Ontology
- C. elegans Phenotype Ontology
- Environment Ontology
- Mouse Pathology Ontology
- Plant Structure Ontology

Influence of the protease flap movement on calculated binding energy

The beta-strand flap is the most flexible region in the HIV-1 protease. It is normally 7 Å RMSD from the active site and is in an open conformation in the native state [22,23]. The protease undergoes significant structural changes on binding to an inhibitor. The two flaps fold over the inhibitor to form a tunnel- shaped active site and are held in this close position by hydrogen bonding from Ile50 and Ile50' NH groups of the enzyme to a water molecule, which in turn is hydrogen bonded to the P2 and P1' CO groups of the inhibitor [24]. The bonding stabilizes the flaps in a closed position and inhibits the activities of the enzyme.

MD simulation has been used to study the movement of the flap region of HIV-1 protease with a ligand [25-30]. The flaps initially opened to an all-atom RMSD of 25 Å within 200 ps and became completely open at the end of a 10 ns simulation. In this study (Figure 2), the flaps opened up and moved away from the x-ray structure from 0.54 Å at 0.1 ps to 3.30 Å RMSD at 10 ps (the flap RMSD was calculated from residue 40 to 60 of each protein chain). These movements, after 0.1 ps of simulation, are inversely correlated with the quality of the binding energy prediction. As shown in Table 1, the correlation coefficient significantly decreased from 0.87 at 0.1 ps to 0.74 at 10 ps as the all-atom flap RMSD increased from 0.54 to 3.30 at 0.1 and 10 ps, respectively.

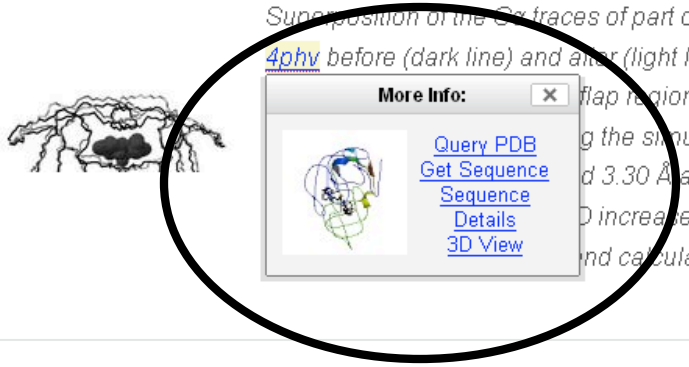


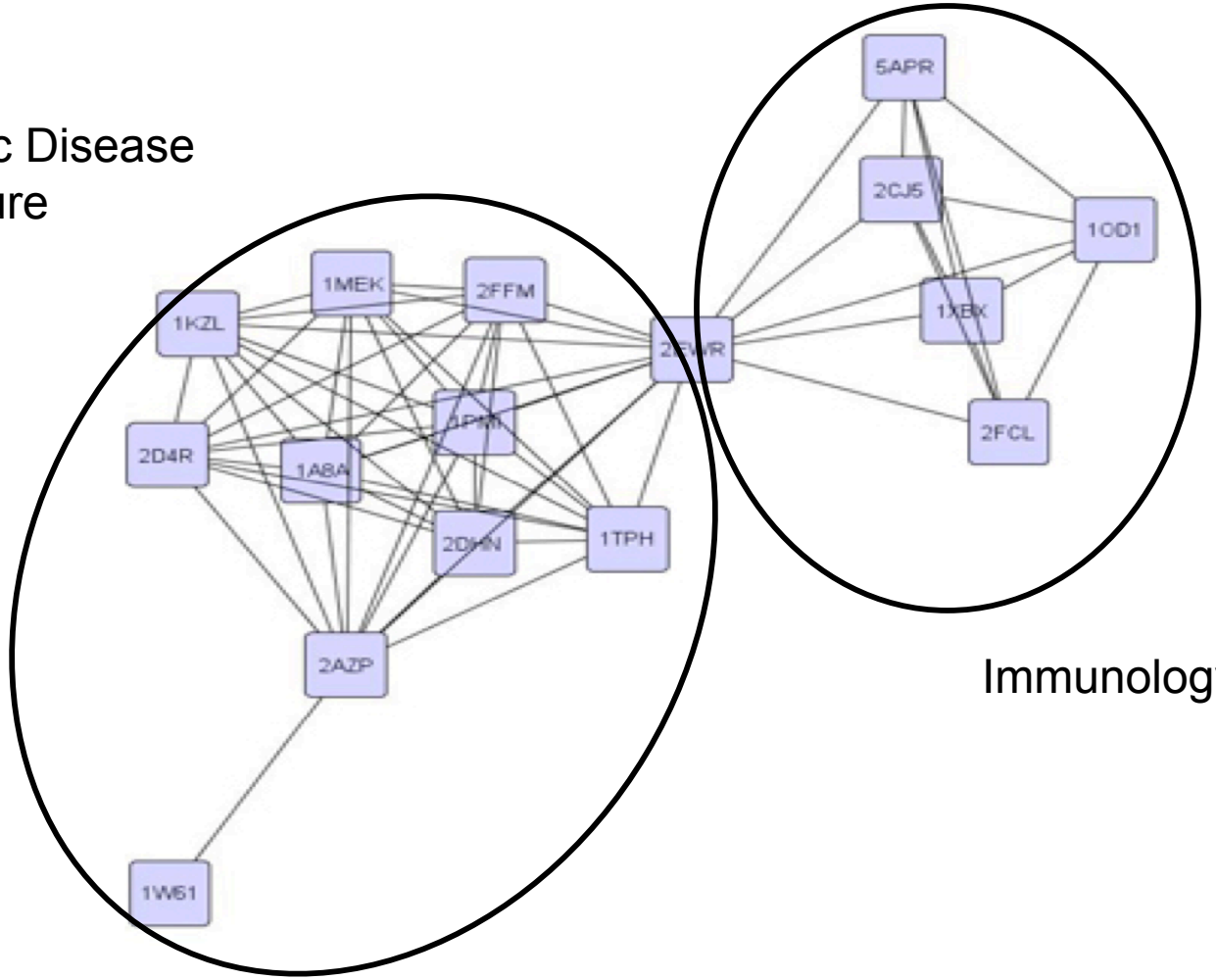
Figure 2

Superposition of the Cα traces of part of the HIV-1 protease x-ray structure 4phv before (dark line) and after (light line) 10 ps of MD simulation bound to ligand. The flap region (above the inhibitor) moved away during the simulation, with all-atom RMSDs of 0.54 Å and 3.30 Å at 0.1 and 10 ps, respectively. Generally, after 10 ps RMSD increased, the correlation coefficient of the predicted and calculated binding energies decreased.

Complementarity between the ligand and the binding site is the basic concept behind ligand binding. This is manifest as steric complementarity, i.e. the shape of the ligand is mirrored in the shape of the binding site, allowing molecular interactions between

Data Clustering via the Literature

Cardiac Disease Literature

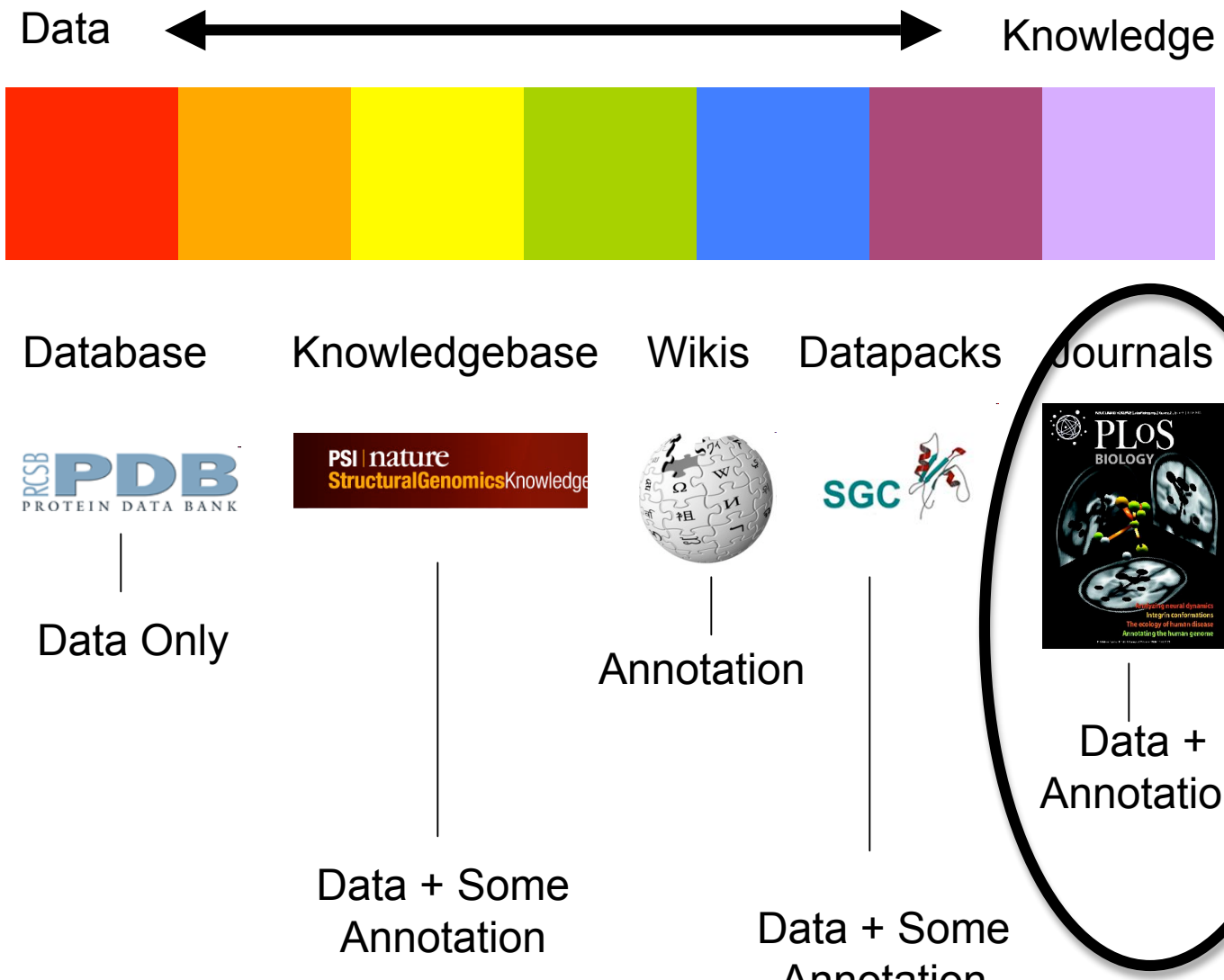


Immunology Literature

We can take this integration of literature and database content a step further…

Enter *PLoS iStructure*

An interactive journal



PLoS iStructure

The Data – Knowledge Spectrum

Enter *PLoS iStructure …*

- Authoring starts with a PDB file
- Annotated molecular views are added – the associated metadata defining those views is stored with the publication
- Other data types can be validated and added through semantic association
- Rich media can be added through embedding
- The resultant publication is peer reviewed and may be published

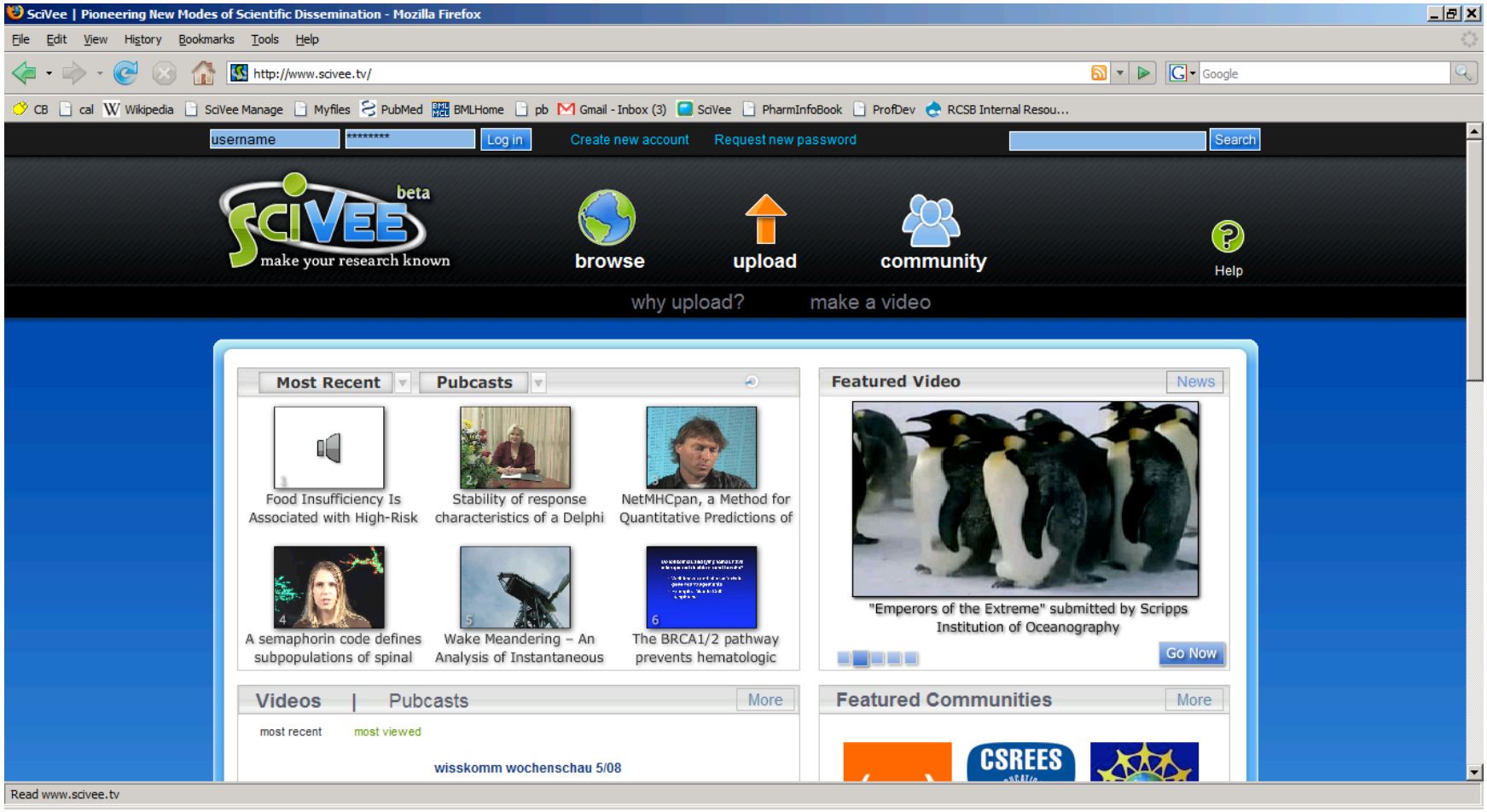
Enter *PLoS iStructure*

- The “reader” has new opportunities for comprehension and analysis
- The journal is an interface to to apply the knowledge found in the paper immediately and seamlessly eg each table is a spreadsheet
- Comparative analysis can be performed directly from the paper

Even a more interactive journal can benefit from rich media types

Video can bring so much to the learning experience

Notion of traditional publications being associated with podcasts and video www.scivee.tv



Pubcast – Video Integrated with the Full Text of the Paper

The opposite view is to embed a SciVee video in a published paper

Provides multiple entry points to that paper and may impact downloads and perhaps citations

In Five Years if There Were More Journals Like iStructure How Would it Impact Curation?

- Certainly it would change the way curation is done. However…
- There is no escapaing the value that a human as a third party can add value to the final product – the product is different is all

Acknowledgements

- BioLit Team
  - Lynn Fink
  - Parker Williams
  - Marco Martinez
  - Rahul Chandran
  - Greg Quinn
- Microsoft Scholarly Communications
  - Pablo Fernicola
  - Lee Dirks
  - Savas Parastitidas
  - Alex Wade
  - Tony Hey

Microsoft

http://biolit.ucsd.edu
http://www.codeplex.com/ucsdbiolit

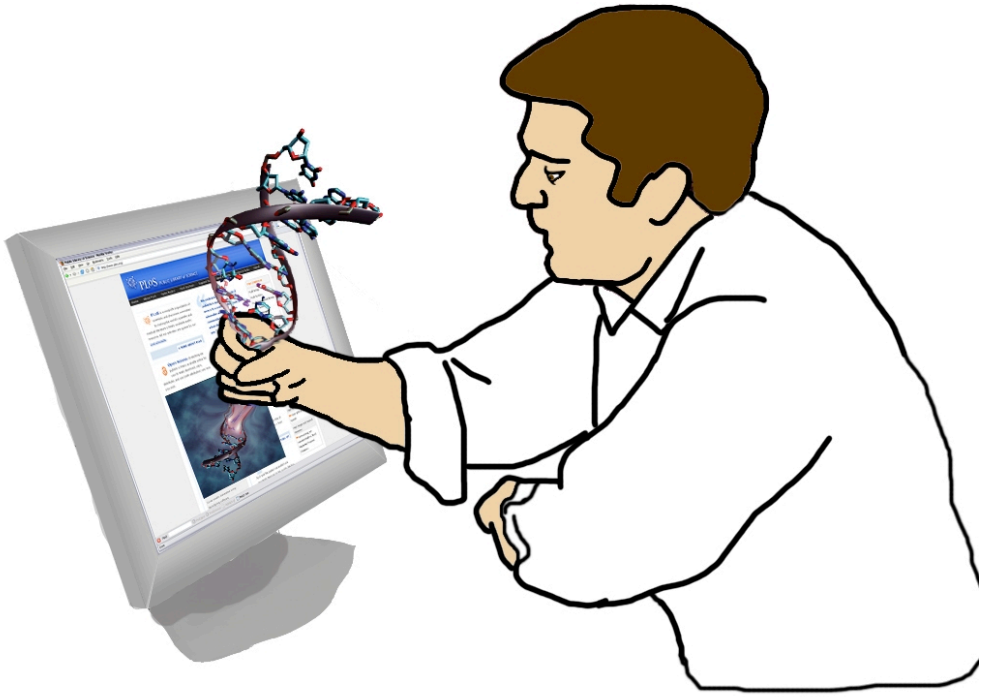
- wwPDB team
- SciVee Team
  - Apryl Bailey
  - Tim Beck
  - Leo Chalupa
  - Lynn Fink
  - Marc Friedman (CEO)
  - Ken Liu
  - Alex Ramos
  - Willy Suwanto

SciVee beta make your research known

http://www.scivee.tv

National Science Foundation
WHERE DISCOVERIES BEGIN

pbourne@ucsd.edu



Questions?

Additional Reading: