Innovation and Openness in Science and the Exceptional Role of the Biological Sciences

Victoria Stodden
Postdoctoral Associate in Law and
Kauffman Fellow in Law and Innovation
Yale Law School

The New Biology:
Pathways to Convergence in the Life Sciences
MIT/Kauffman Seminar
April 7, 2010

Three Broad Groupings of Biological Sciences

- Genomics, bioinformatics, ...
- Pharma, drug and device research, ...
- Hospitals, docs, clinical trials, medical records. ...

Different worlds with different attendant pressures, incentives, and norms, particularly with respect to information sharing.

The 1996 Bermuda Agreement

Primary Genomic Sequence Should be in the Public Domain
It was agreed that all human genomic sequence information, generated by centres funded for large-scale human sequencing, should be freely available and in the public domain in order to encourage research and development and to maximise its benefit to society.

Primary Genomic Sequence Should be Rapidly Released

- Sequence assemblies should be released as soon as possible; in some centres, assemblies of greater than 1 Kb would be released automatically on a daily basis.
- Finished annotated sequence should be submitted immediately to the public databases.

Reaffirmed and Extended Through 2009

- Bermuda 1997 and 1998 (addresses standards and error rates, extension to other organisms);
- Fort Lauderdale 2003 (extension to all sequence data; established "community resource projects" for large scale data production);
- Amsterdam 2008 (extension to proteomics data);
- Toronto 2009 (pre-publication data release; extension to other datasets in biology and medicine that have broad utility).

Result: Massive data availability creating a platform for widespread analysis and discovery

PARASITOLOGY

Key Malaria Parasite Likely Evolved From Chimp Version

For centuries, the origin of the main parasite that causes malaria in humans has remained murkier than the stagnant water loved by the mosquitoes that transmit the killer pathogen. Now an international research team has uncovered compelling evidence that the parasite, *Plasmodium falciparum*, evolved from a relative called *P. reichenowi* that infects chimpanzees. The data support a provocative theory that as human red blood cells evolved a way to dodge *P. reichenowi*, they became highly vulnerable to *P. falciparum*.

forward and places *P. falciparum* clearly in a cluster of chimp parasite species," says parasitologist Julian Rayner of the Wellcome Trust Sanger Institute in Hinxton, U.K.

In 1920, German researcher Eduard Reichenow claimed to have found *P. falciparum*—which in people causes what researchers call "malignant malaria"—in chimpanzees. But attempts to infect chimps with lab isolates failed, which led two British researchers, Donald Blacklock and Saul Adler, to inject themselves a few years later with blood taken from a *Plasmodium*-infected chimpanzee. Neither of them became ill, and they concluded that chimps had a different parasite, which they named *reichenowi*.

In the 1990s, Francisco Ayala, a co-author of the new paper and the former Ph.D. adviser of Rich, revived interest in the long-ignored chimp parasite. Not persuaded by a study that linked the human parasite's evolution to a *Plasmodium* found in birds, Ayala noticed that

ScienceNOW.org

From *Science*'s Online Daily News Site

Gorilla Virus in Our Midst
Researchers are shaking up the HIV fami
tree again. For the first time, investigato
have found what looks like a gorilla ver-
sion of the AIDS virus in a person. They d
not know how the woman became infect
but suspect that other humans harbor a
similar virus. The possibility that gorillas
can transmit the virus to humans further
underscores the danger of butchering th
apes or keeping them as pets, which still
occurs in some African communities.
http://bit.ly/Io0Q7

Dinosaur Study Backs Controversial Fi
When scientists reported 2 years ago that
they had discovered intact protein frag-
ments from a 68-million-year-old *Tyran-
nosaurus rex*, the skeptics pounced. They
argued that one of the main lines of evi-
dence, signatures of the protein fragmen
taken by mass spectrometry, was flawed.
But now a reanalysis of that mass-spec da

Opening of Data in Other Areas Beyond Genomics

For example,

- Patient-level medical records (hospitals, docs),
- Patient-generated data ie. patientslikeme.com, lybba.org, ...

Especially fuels drug and device discovery and development.

Open data issues generally are bringing to the surface complex research ownership issues, and deep ethical concerns (ie. biobricks and DIYbio).

From Computational Science:

The Reproducible Research Standard (Stodden 2009) addresses transparency and reproducibility issues in computational science:

- to realign IP rights with scientific norms,
- reinstate reproducibility in computational science,
- release of scientific research, including code and data, for verification, wide reuse, development...

To satisfy the Reproducible Research Standard, use attribution-only licensing on text, figures, code, and certify data in the public domain.

Legal Nuts and Bolts for Open Reproducible Science

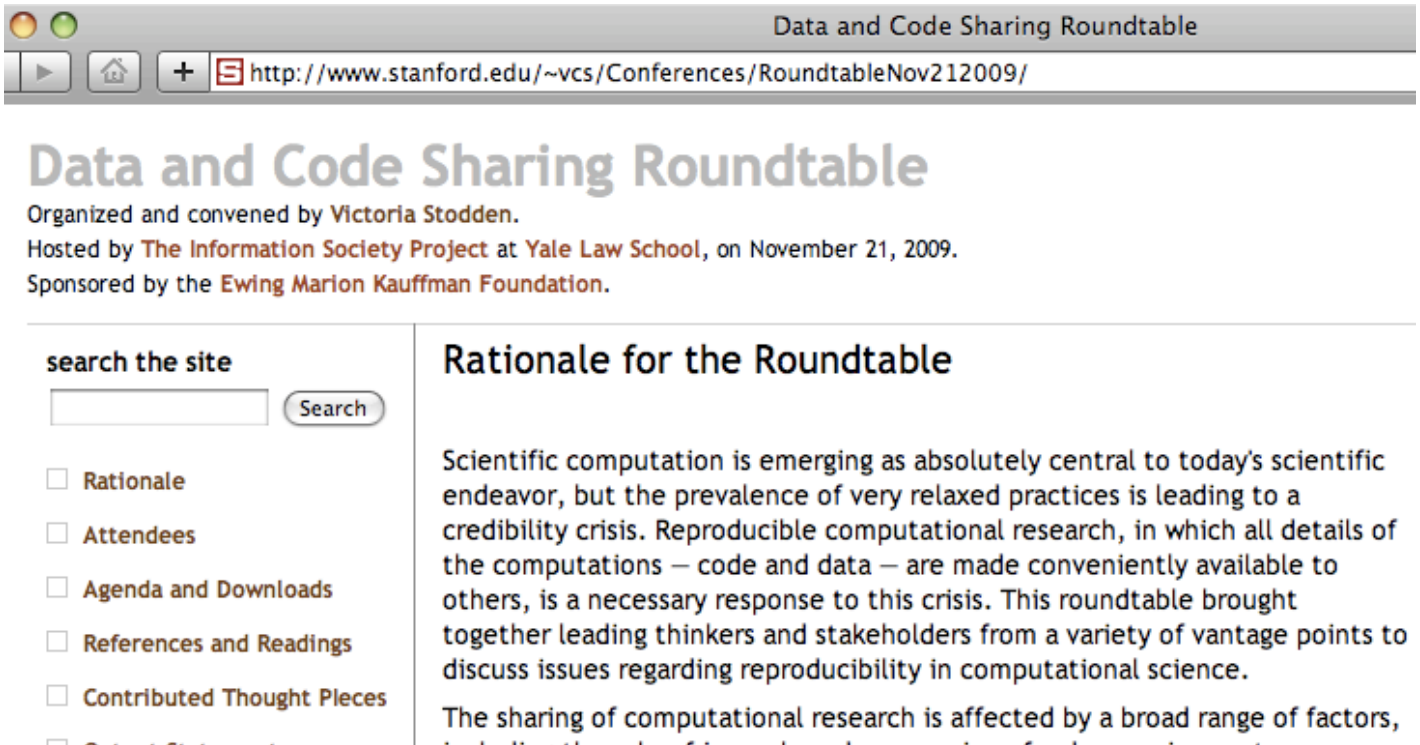
1. Creative Commons Attribution license (CC BY) on media such as text, figures,
2. Attribution license on code: such as Apache 2.0, MIT, LGPL,
3. Data under CC0 or Science Commons Open Access Data Protocol,
4. "original selection and arrangement" of the data, under CC BY or attribution open source license.

Many more ownership details to be sorted out in various cases (e.g. confidentiality issues, proprietary data/code..).

Data and Code Sharing Roundtable, November 21 2009

Gathered 32 folks at Yale Law School: prominent scholars, university administrators, funders, and journal representatives.
Produced (almost):

1. Thought pieces published on the web,
2. Position statement as output declaration.



Engineering-oriented Biological Research: Example

Not only techniques developed in computer science but integration of many different data sources.

COCEBI-649; NO OF PAGES 12 ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Cell Biology

ELSEVIER

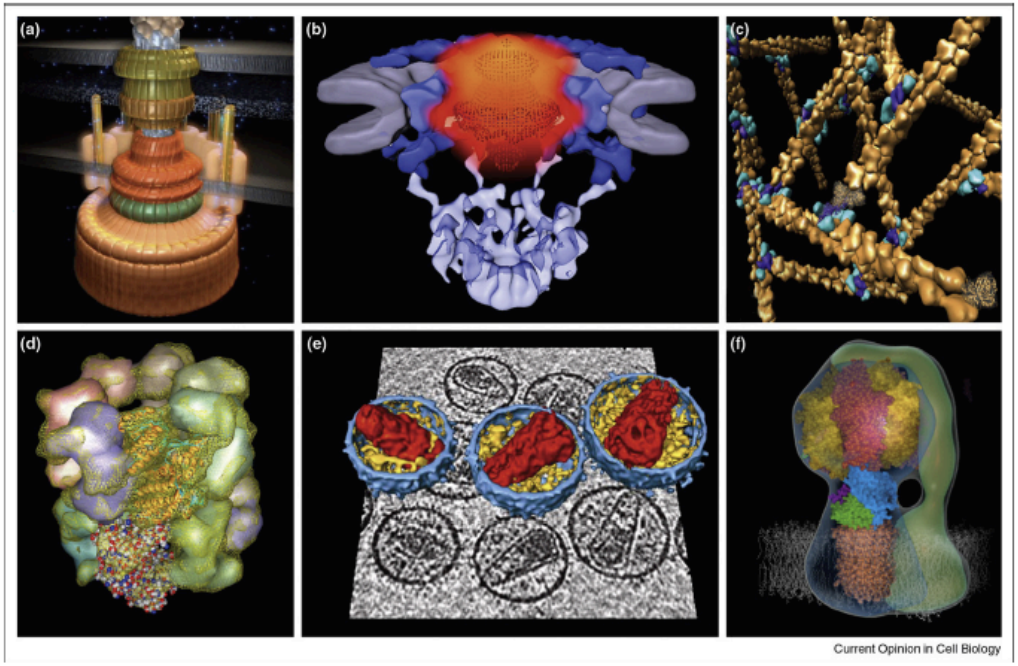
The structural dynamics of macromolecular processes

Daniel Russel[1], Keren Lasker[1,2], Jeremy Phillips[1,3], Dina Schneidman-Duhovny[1], Javier A Velázquez-Muriel[1] and Andrej Sali[1]

Dynamic processes involving macromolecular complexes are essential to cell function. These processes take place over a wide variety of length scales from nanometers to micrometers, and over time scales from nanoseconds to minutes. As a result, information from a variety of different experimental and computational approaches is required. We review the relevant sources of information and introduce a framework for integrating the data to produce representations of dynamic processes.

No single technique, computational or experimental, is able to span all relevant spatial and temporal scales (Figure 3). For static complexes, for example, X-ray crystallography can generate atomic structures of the components, while single particle cryo-electron microscopy (cryo-EM) can provide average mass density maps of the whole assembly at nanometer resolution for the whole assembly. For processes, computer simulations are beginning to reach the microsecond time scale, while

Dynamic Macromolecular Processes



(a) Locomotion of a cell is enabled by a reversible rotary propeller of the bacterial flagellum. (b) Nucleocytoplasmic transport of macromolecules occurs in a regulated fashion through the nuclear pore complex. (c) A number of cellular functions, including muscle contraction, cell motility, cell division, and cytokinesis, depend on the assembly and maintenance of branched actin filaments. (d) The folding of many proteins is catalyzed inside the chaperonin cavity. (e) The HIV-1 core assembles inside the maturing virion. (f) Synthesis of ATP in mitochondria and

Emergence of First Principles

- move to replicable science (bio community most advanced in the establishment of data repositories, journals with code/data submission, ie *Biostatistics*, microarray publications, PDB,...)
- current research in "building blocks of life" - computationally intensive attempt at understanding necessary components for life (Venter),
- many others...

Key ingredients in acceleration of convergence

To promote the adoption and development of tools, techniques, and the application of knowledge:

- open innovation systems (open code vital, not just open data):
  - permits interdisciplinary collaboration and sharing,
  - facilitates the import of these research tools and technologies,
- many other fields are are adopting various engineering and computational techniques (e.g. climate research, social sciences),
- new data are creating platforms for analysis: medical records, doc experience, knowledge pooling,
- maximize discovery and scientific integrity through openness and reproducibility.

Conclusions

- Massive computation revolutionizing scientific research, with strongest impact in (and greatest leadership from) the biological sciences.
- Must be accompanied by methodology and platform for openness and reproducibility in scientific research.

References:

- "Enabling Reproducible Research: Open Licensing for Scientific Innovation"
- "15 Years of Reproducible Research in Computational Harmonic Analysis"
- "The Legal Framework for Reproducible Research in the Sciences: Licensing and Copyright,"

`http://www.stanford.edu/~vcs`