Applications for Open Source in Proteomics

Workshop on Strategies for Improving Reliability in Protein/Peptide Identification

Organized by NCI and NIST

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

- Philip Andrews
  - National Resource for Pathways and Proteomics (NRPP)
  - Michigan Proteome Consortium

Target Points

- Why OS is important for technology development
- Current state of Open Source in Proteomics(very brief and limited)
- The relationship between Open Source and Open Access to data
- Some thoughts on OS and Proteomics

CENTRAL PROTEOMIC TECHNOLOGIES FOR CANCER

Challenges in Proteome Informatics

- **Proteome technologies evolve rapidly.**
  - Software always lags behind hardware.
  - Software always lags behind applications.
- **Proteomics deals with large and often complex data sets.**
- **Proteomics cannot be done without investment in computational infrastructure.**

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

For These Reasons……

- Most technology development in proteomics requires development or modification of computational tools.
- Instrument development is often negatively impacted by software development (cost, development time, functionality, ……).
- If the perceived market for a new product is too small, resources will not be allocated to software support.

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Advances in Proteomics Are Constrained by Software and Data Management.

- Application to biological problems.
- Technology development (instruments, chemistries, methodologies).

Partial Solutions

- Put more resources into Proteome Informatics.
  - Approximately one third of any proteomics budget should be bioinformatics/computational.
  - Not going to happen, in general.
- Build an open source community and support it.

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

What is Open Source?

open-source

adj : of or relating to or being computer software for which the source code is freely available."

WordNet (r) 2.0

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

- *It is not necessarily free software.*

Licensing Issues

- Are dependent on:
  - Developer policies
  - Institution policies
  - Funding agencies
  - Existing IP license arrangements
  - IP landscape
- Our preferred
  - Formerly Apache (liberal) license
  - Currently Creative Commons license

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Large OS Projects Can Be Successful and Widely Used

- Linux
- Apache web server.
- BIND provides DNS.
- Firefox, Thunderbird
- Open Office
- OpenSSL

But you don't need massive coding projects for OS to be useful.

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Selected OS Proteomics Projects

- ISB (Eng, Nesvizhskii- Peptide/Protein Prophet, SpectraST, etc.)
- Beavis (X!Tandem, X!Hunter, etc.)
- Vanderbilt CPTAC (Myrimatch, etc.)
- NRPP (Tranche, Bonanza, MSExpedite, PRIME, etc.)
- FHCRC (CPAS, etc.)
- Pacific NW Labs (AMT software, etc.)
- UCSD (Pevzner-PepIdent, etc.)
- NCBI (Geer-OMSSA)
- Many others (www.proteomecommons.org and dedicated sites)

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

How Is Open Source Software Useful for Proteomics?

- Allows interpretation of data from new technologies.
- Enable new approaches to proteomics.
- Allows multiple uses of datasets.
- Provides documentation for new algorithms.
- Provides tool sets for developers and researchers.
- Apps represent useful examples or protocols.

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

GPM Usage

Slide from Ron Beavis

Synergies Between OS & Commercial Tools

- Some natural progression from Open Source tools to commercial packages as field matures.
- OS tools can inform commercial software development- directly or indirectly
- Some OS software can be used directly by commercial software (X!tandem/Scaffold)
- Some OS software will make transition to commercial products

CLINICAL PROTEOMIC TECHNOLOGIES for CANCER

Open Source Modules/Tools Needed in Proteomics

- MS/MS search tools
- PMF search tools
- De novo search tools
- Data evaluation tools (mass spectra and search results)
- Signal processing/peak detection, deisotoping, etc.
- Data viewers
- Higher order tools for curation and analysis
- Data management

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

OS Summary: A (partial) Solution: for Proteomics

- Progress in Proteomics will be faster if we have a robust open source community.
- Open source efforts allow the community to respond to new technologies rapidly.
  - Software always lags behind hardware.
  - Software always lags behind applications.
- Open source allows the community to respond to its own needs.
- Cost of development is shared.
- Open Source is compatible with commercial proprietary software.

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Key Concepts that Can Benefit Proteomics

- Code and data should be readily available and free to use, but credit should always be given to the authors.
  - Free includes both commercial and non-commercial use cases
  - Acknowledging credit is analogous to citing publications.
  - The goal is to provide a setting that allows existing work to both be used creatively and publicly critiqued.

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

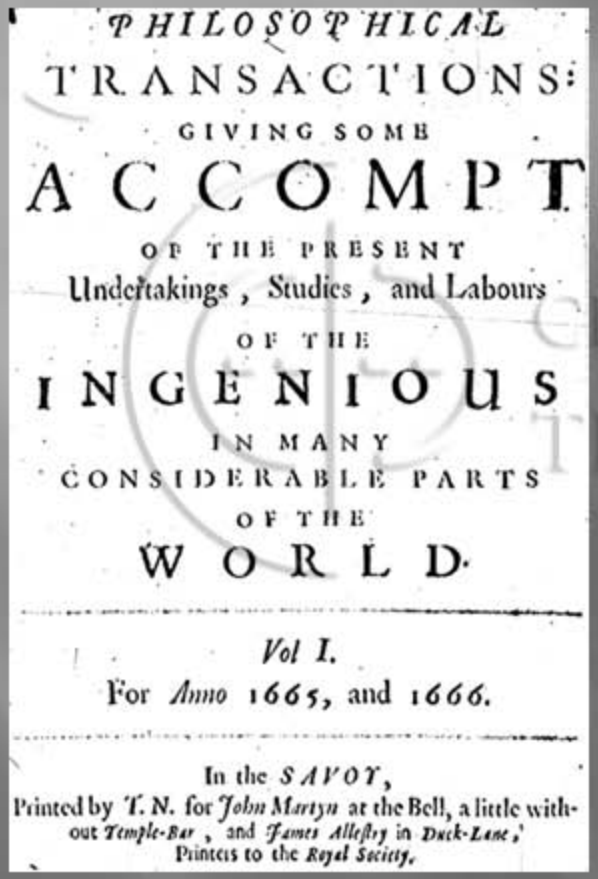
Open-Source for Proteomics

- Code must be publicly available and free (technically, "free software")
  - Use appropriate free, open-source license
    - http://www.fsf.org/
  - Formal releases and an open repository for getting the latest, unreleased code
    - Subversion (http://subversion.tigris.org/)
  - Available without restriction via the WWW
- Code also means relevant data
  - Both peak lists and spectra, gel images...

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

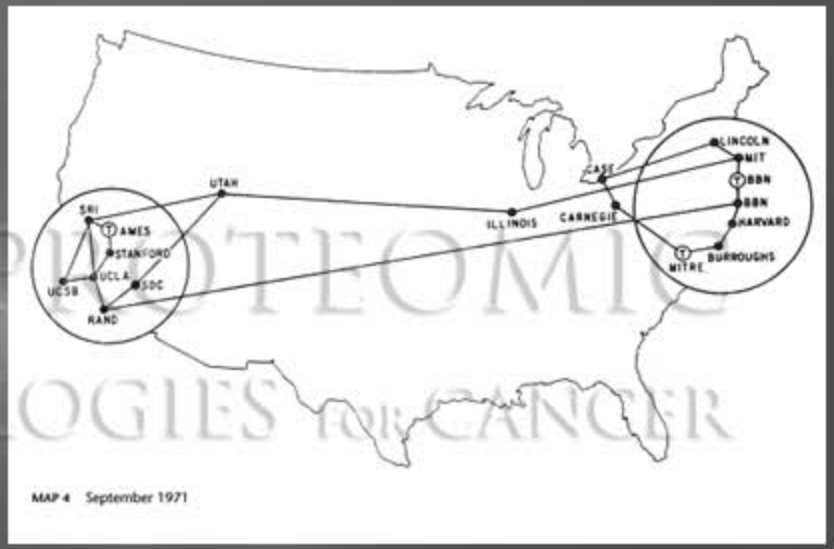
An indispensable driving force for science: *Communication*

Research thrives on the ability of researchers to communicate


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.



Arpanet, 1971

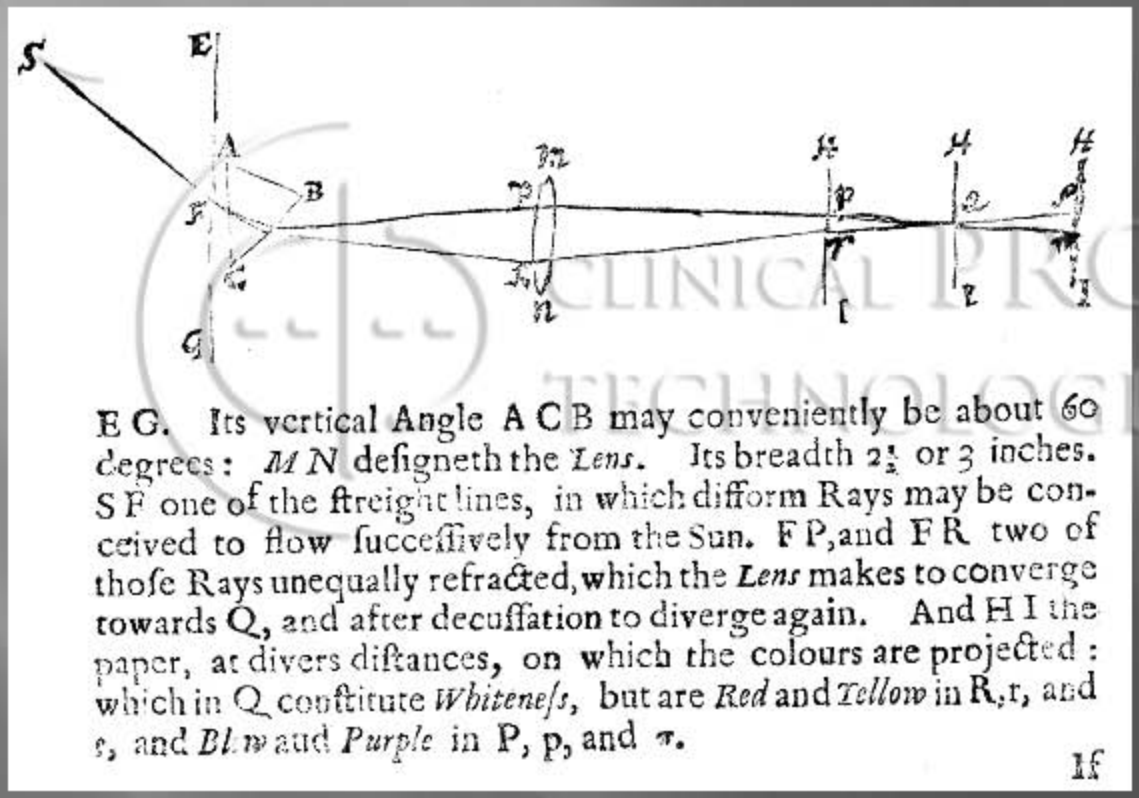
Philosophical Transactions
(1665)
vol 1, no 1

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Images: wikipedia

The point of an article is to present data analysis

Should an article represent, reference or incorporate its data?



E G. Its vertical Angle A C B may conveniently be about 60 degrees: *M N* designeth the *Lens*. Its breadth 2½ or 3 inches. S F one of the ſtreight lines, in which difform Rays may be conceived to flow ſucceſſively from the Sun. F P, and F R two of thoſe Rays unequally refracted, which the *Lens* makes to converge towards Q, and after decuſſation to diverge again. And H I the paper, at divers diſtances, on which the colours are projected: which in Q conſtitute *Whiteneſs*, but are *Red* and *Yellow* in R, r, and ſ, and *Blue* and *Purple* in P, p, and π.

If

Newton, Isaac. Theory about Light and Colours. *Philosop. Trans.* 6:3075-3087 (1667).

Article format has remained virtually unchanged in the last 300 years .

Should we care?

Yes, because the original goal of repeatability is still important.
Proteomics is still mainly data-centric.

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

What is Open Access?

- Free and public access to digital scientific and scholarly information.
  - Enabled by the Internet
    - PubMed Central
    - Public Library of Science
    - Creative Commons

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

How Is Open Access to Datasets Useful?

- **Datasets are required for validation of algorithms and software** (Keller, Gygi, Aurum, etc.).
- **Re-mining (reuse) of datasets** (genome annotation, biomarker discovery, systems biology, spectral libraries etc.)
- **Validation of interpretations**.

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Proteomics Databases and Repositories

- Peptide Atlas (www.peptideatlas.org)
- The GPMdb (www.thegpm.org)
- PRIDE (www.ebi.ac.uk/**pride**/)
- HPRD (www.hprd.org)

- Tranche (www.proteomecommons.org/dev/dfs)

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Tranche Is a Secure Transactional System Built on a P2P Model.

- Secure public or private data dissemination
- Provides a proper citation (stable, unique)
- Provides data provenance and integrity
- Independent of file format
- Allows forward annotation, file retraction, and related capabilities
- >5,000 projects online
- Linked to centralized databases (PRIDE, HPRD)

CLINICAL PROTEOMIC TECHNOLOGIES for CANCER

Disjunctions Between Printed Articles and the Information Century

As journals stick to print-centric online formats)…..

- Larger, more complicated data generated in most fields.
  - *Hard to reduce to print form (and supplemental information is stopgap).*
- Harder to review properly.
  - *Validation /peer review incomplete or absent.*
- Potential data loss over time.
  - *Researchers move, institutions change data storage policies, migration not economic, software /formats no longer supported.*
- Increasing loss of the original purpose – *reproduction of results and extension of public knowledge.*

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

References and Additional Links

- The Cathedral and the Bazaar, etc. *Eric S. Raymond*
- *www.creativecommons.org*
- *The Political Economy of Open Source Software*, BRIE Working Paper 140, Steven Weber
- http://brie.berkeley.edu/~briewww/pubs/pubs/wp/wp140.pdf
- The World is Flat *Thomas L. Friedman*

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER

Acknowledgements

- Jayson Falkner
- Mustafa Ünlü
- Panagiotis (Takis) Papoulias
- Peter Ulintz
- AlexeyNesvizhskii
- All our Tranche collaborators (tranche.proteomecommons.org)

- NCRR P41: for development of Tranche, PRIME, Bonanza, MSExpedite, IOFramework, etc.)
- NCI CPTAC: Partial ongoing support for Tranche.

CLINICAL PROTEOMIC TECHNOLOGIES FOR CANCER