What do academic researchers want and how can this facilitate blue skies research?

John Quackenbush

CASIMIR Workshop

20 May 2009

The Computational Biology and Functional Genomics Laboratory

at the Dana-Farber Cancer Institute and Harvard School of Public Health

Genomic science is wonderful in that it brings together representatives of so many disciplines – clinicians, bench biologists, statisticians, bioinformatics scientists – all of whom tend to consider the others intellectual peasants.

– Isaac “Zak” Kohane

Levels of Biological Information

'omics

DNA — Genomics

mRNA — Functional Genomics

Proteins — Proteomics

Informational M[illegible]thy[illegible] — Metabolomics

Informational N[illegible] — [illegible]ns Biology

Cells — [illegible]llular Biology

Organs — Medicine

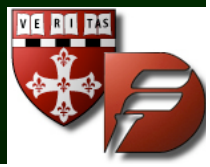
Individuals — Medicine

Populations — Genetics

Ecologies — Ecology

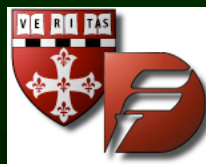
The Future!

Traditional Biology



Assumptions: Key Elements in Biomedical Research

- Increasingly, research is driven by access to relevant, well characterized, well annotated samples and models
- New technologies such as DNA microarrays, next generation sequencing approaches, proteomics, metabolomics, etc. are driving discovery
- Integration of genomics with phenotype and the analysis of diverse biological samples can lead to significant advances
- Progress requires the development of both biological and data resources as well as the creation of new tools to use these to advantage.



Needs and Challenges

A fully integrated resource of relevant samples indexed and linked to phenotypic information

- There is a need for easy-to-use and standardized forms for collecting phenotype data and entering these into a database
- This has to include methods for effectively capturing ‘omics data (see next slide)
- Standardized methods for describing phenotype (ontologies)
  - a rose by any other name is a rose, you just can’t find it in the database
- Sample tracking and distribution mechanisms that provide secure access to relevant data on models
- The extension of the concept of “orthology” from gene to pathway to phenotype

Needs and Challenges II

Genomic, microarray, proteomic, metabolomic data collected using standardized protocols and stored in a standardized format

- Standardized protocols for collection of samples
- Standardized protocols for collection of ‘omics data
- The ability to adapt to changing protocols ($1000 genome)
- Standards for data reporting (ala MIAME)
- Ontologies for phenotypes (again)
- Standards and ontologies are only useful if there are tools that implement them
- Quality standards for assessing assays and studies
- A means to recapture data from the models
  (if we give you a sample, you return the data)

Needs and Challenges III

A central database providing access to the relevant data in an intuitive fashion

- Database interoperability has been a dream for 20 years and will remain so
- The finished genome is not
- The creation of such a database is limited by sample tracking (FedEx) and data collection
- A database will not be used unless there is an emphasis on developing usable interfaces that provide access to raw data and provide information that answer common questions investigators have
- This is not sexy science

Needs and Challenges IV

Tools for data integration and interpretation that provide access to laboratory scientists and clinicians

- Bioinformatics research is often divorced from the analytical needs of front-line biologists
- Developing software tools has to happen in close partnership between laboratory and computational scientists and bioinformatics developers
- Many methods are developed, few are chosen, but we need to know when to make the transition
- Data analysis in the absence of biology is not a useful exercise
- The genome sequence can serve as a great organizer

The Fallacies

- Genomics has taught us how to handle large-scale ‘omics data
- The Genome is finished
  - SNPs
  - Large-scale polymorphisms
  - A complete catalogue of genes
  - Pathways and networks
  - Orthologies
- Linking genes to proteins and metabolites is a simple task
- Researchers need cutting-edge bioinformatics tools
- Once a database is built, it is finished
- Bioinformatics is cheap
- Data collection is expensive

The Dream

- A fully integrated resource of models indexed and linked to relevant ‘omic and phenotype data
  - appropriate tissue samples
  - blood
  - urine
- Genomic, microarray, proteomic, metabolomic data collected using standardized protocols and stored in a standardized format
- A central database providing access to the relevant data in an intuitive fashion
- Tools for data integration and interpretation that provide access to laboratory scientists

Lessons Learned

- 'omics data is a good deal more complex than genome sequence data because we need to capture ancillary data
- Standards cannot be imposed top-down, but should be built with community involvement
- Creating standards without (freely accessible, open source) tools to implement them is pointless
- Standards must be developed in collaboration with data repositories, data generators, and instrument manufacturers
- Standards must evolve as our understanding grows
- One needs carrots and sticks – from funders and journals – to make standards work
- Be careful what you wish for

The best way to predict the future is to invent it.

-Alan Kay, inventor (1940-)

The future is here. It's just not widely distributed yet.

- William Gibson