The Measurement Challenge in Bioscience – Dealing with the Data

National Laboratory Association
Sept 2009

E Jane Morris,
African Centre for Gene Technologies
The research paradigm in bioscience is changing -

- **Old**
  - Hypothesis driven
  - Experiments designed to answer a specific question
  - Measuring individual variables

- **New**
  - Data driven
  - Large scale projects
  - Look at a biological system holistically
  - Measuring multiple variables
From Sequence to Function

- **Genome sequencing**
  - Around 30,000 genes in the human genome
  - Less than 2% of the genome codes for proteins

- **Functional genomics**
  - How do genes function and how are they regulated

- **Transcriptomics**
  - Study of genes expressed through mRNA

- **Comparative genomics**
  - Comparison of DNA between organisms

- **Structural genomics**
  - 3-D structures of proteins from all protein families

- **Proteomics**
  - Study of all the expressed proteins

- **Metabolomics**
  - Study of all the metabolites in an organism
From genomics to X-omics

- Genome is virtually static: roughly well-defined for an organism
- Gene transcription is in response to metabolic state and external environment
- Transcriptome, proteome, metabolome etc continually change in response to external and internal events!

There is an unlimited number of cellular states of a given cell or organism!
One genotype – two phenotypes!

Tadpoles

Bug-Eyed Tree Frog
Measuring biological systems

Measuring can be approached from many scientific perspectives

- Computer scientist
- Biochemist
- Geneticist
- Physiologist
- Cell biologist
- Structural biologist
A world of interlinked data

Biologically-delineated views of the world
A: plant biology
B: epidemiology
C: microbiology
...and...

Generic features (‘common core’)
- Description of source biomaterial
- Experimental design components

Technologically-delineated views of the world
A: transcriptomics
B: proteomics
C: metabolomics
...and...

Diagram copied from presentation by Chris Taylor, EMBL-EBI & NEBC
...or viewed slightly differently

| Investigation: | Medical syndrome, environmental effect, etc. |
| Study:         | Toxicology, environmental science, etc.      |
| Assay:         | Omics and miscellaneous techniques          |

Diagram copied from presentation by Chris Taylor, EMBL-EBI & NEBC
How can bioscientists cope if there are no standards for measurement, data reporting and data integration?
Some data challenges

- Determining the quality of data
- Standardizing the data
- Storing the data
- Transmitting and sharing the data
- Integrating data sets and data types
- Making sense of the data in the context of cell function
- Understanding networks and interactions
Data’s Shameful Neglect

“Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly.”

“More and more often these days, a research project's success is measured …by the data it makes available to the wider community. Pioneering archives such as GenBank have demonstrated just how powerful such legacy data sets can be for generating new discoveries - especially when data are combined from many laboratories and analysed in ways that the original researchers could not have anticipated.”

“All but a handful of disciplines still lack the technical, institutional and cultural frameworks required to support such open data access”

Nature Editorial 10 September 2009
The sheer size of datasets generates its own problems

- Human genome – 3 Gb data
- 1 Genome analyzer can already generate 2.5Gb per day – and growing exponentially!
- And don’t forget the multiple transcriptomes proteomes etc
- Yale Center for High Throughput Biology estimates that it will generate around 50 terabytes of data each year—enough to fill more than 100 desktop computers.
Some light on the horizon!

MIBBI – Minimum Information for Biological and Biophysical Investigations

- Many “Minimum Information” (MI) checklists being developed for experimentation and data reporting
- The checklists and standards need to allow for integration of the data!
Some MIBBI projects

- **CIMR** Core Information for Metabolomics Reporting
- **MIABE** Minimal Information About a Bioactive Entity
- **MIACA** Minimal Information About a Cellular Assay
- **MIAME** Minimum Information About a Microarray Experiment
- **MIAPA** Minimum Information About a Phylogenetic Analysis
- **MIAPAR** Minimum Information About a Protein Affinity Reagent
- **MIAPE** Minimum Information About a Proteomics Experiment
- **MIARE** Minimum Information About a RNAi Experiment
- **MIASE** Minimum Information About a Simulation Experiment
- **MIENS** Minimum Information about an ENvironmental Sequence
- **MIFlowCyt** Minimum Information for a Flow Cytometry Experiment
- **MIGen** Minimum Information about a Genotyping Experiment
- **MIGS** Minimum Information about a Genome Sequence
Some MIBBI projects

- **MINSEQE** Minimum Information about a high-throughput Sequencing Experiment
- **MIPFE** Minimal Information for Protein Functional Evaluation
- **MIQAS** Minimal Information for QTLs and Association Studies
- **MIqPCR** Minimum Information about a quantitative Polymerase Chain Reaction experiment
- **MIRIAM** Minimal Information Required In the Annotation of biochemical Models
- **MISFISHIE** Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments
- **STRENDA** Standards for Reporting Enzymology Data
- **TBC** Tox Biology Checklist
- **MIMIx** Minimum Information about a Molecular Interaction Experiment
- **MIMPP** Minimal Information for Mouse Phenotyping Procedures
- **MINI** Minimum Information about a Neuroscience Investigation
- **MINIMESS** Minimal Metagenome Sequence Analysis Stan
## Comparison of MIBBI-registered projects [21]

### Version 0.7  (2008-04-10)

<table>
<thead>
<tr>
<th>CONCEPT</th>
<th>SPECIALISATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>study inputs</td>
<td>study design</td>
</tr>
<tr>
<td></td>
<td>generic organism</td>
</tr>
<tr>
<td></td>
<td>cells / microbes</td>
</tr>
<tr>
<td></td>
<td>plant</td>
</tr>
<tr>
<td></td>
<td>animal</td>
</tr>
<tr>
<td></td>
<td>mouse</td>
</tr>
<tr>
<td></td>
<td>human</td>
</tr>
<tr>
<td></td>
<td>population</td>
</tr>
<tr>
<td></td>
<td>environmental sample</td>
</tr>
<tr>
<td></td>
<td>environment / habitat</td>
</tr>
<tr>
<td></td>
<td>in silico model</td>
</tr>
<tr>
<td>study procedures</td>
<td>organism maintenance</td>
</tr>
<tr>
<td></td>
<td>animal husbandry</td>
</tr>
<tr>
<td></td>
<td>cell / microbe culture</td>
</tr>
<tr>
<td></td>
<td>plant cultivation</td>
</tr>
<tr>
<td></td>
<td>acclimation</td>
</tr>
<tr>
<td></td>
<td>preconditioning / pretreatment</td>
</tr>
<tr>
<td></td>
<td>organism manipulation</td>
</tr>
<tr>
<td>assay inputs</td>
<td>generic study input</td>
</tr>
<tr>
<td></td>
<td>organism part</td>
</tr>
<tr>
<td></td>
<td>organism state</td>
</tr>
<tr>
<td></td>
<td>organism trait</td>
</tr>
<tr>
<td></td>
<td>biomolecule</td>
</tr>
<tr>
<td></td>
<td>synthetic analyte</td>
</tr>
<tr>
<td></td>
<td>silencing RNA reagent</td>
</tr>
</tbody>
</table>

### Granularity

<table>
<thead>
<tr>
<th>Coarse</th>
<th>Medium</th>
<th>Fine</th>
</tr>
</thead>
</table>

### Maturity

- Planned
- Drafting
- Release

[1] Denotes that a specification is provided as a suite of related documents.
The MIBBI Project (www.mibbi.org)

Diagram copied from presentation by Chris Taylor, EMBL-EBI & NEBC

Interaction graph for projects (line thickness & colour saturation show similarity)
The end goal -

- Analyse and integrate the data to build robust models of biological processes!