“MGED Standard implementation: establishing an infrastructure for sharing microarray data”

Philippe Rocca-Serra
Scientific Data Manager

Microarray Informatics Team @ EMBL-EBI
European Bioinformatics Institute, Cambridge, UK

NETTAB meeting, Bologna, 27-28th November 2003
Talk structure

- Standardization efforts
  - Microarray standards by the MGED Society

- Infrastructure at the EBI
  - ArrayExpress Functional implementation of MGED standards
  - ExpressionProfiler

- Future and Development
  - Infrastructure
  - Standards
International network of 5 research institutes dedicated to research and service in molecular biology.

European Molecular Biology Laboratory

- Heidelberg
- Grenoble
- Hamburg
- Monterotondo

Services
Building, maintaining and making available databases

Research in bioinformatics and computational molecular biology

Industry
Promote standards

EBI

Service
Research
Training
Industry

Microarray informatics Team
Standards are important…..

This is not a recent issue…..

Mensa pondereria (Pompeii)
Why are standards so important?

Database or LIMS

Experiment design

Information collection

Data analysis

Results

www.myresults.org
Microarray data.....inflation

1st paper by Stanford University

Less then 10 papers

Data deluge -> Over 1000 papers!

Alvis Brazma @ EBI Industry Programme
‘Promoter prediction from co-expressed genes’

- Large datasets
- Different platforms, surface types
- Different data formats
- Different level of details
- No infrastructure, no public database!
Storing DNA-Microarray generated information

Key questions….

- What should be the information to be recorded?
- How to make sets of information comparable?
- How to record the information?
- Where to store the information?
- How to retrieve and display information?
To address these issues....The MGED response

MGED = Microarray Gene Expression Data Society

<table>
<thead>
<tr>
<th>EBI + Academics, e.g.</th>
<th>Companies e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIGR</td>
<td>Affymetrix</td>
</tr>
<tr>
<td>NCBI</td>
<td>Agilent / Rosetta</td>
</tr>
<tr>
<td>U Penn.</td>
<td>lobion</td>
</tr>
</tbody>
</table>

- **MIAME - WG**
  - A.Brazma
  - EBI
  - Standard for contextual information

- **MAGE - WG**
  - P.Spellman
  - U. California
  - Standard for data representation and exchange

- **Ontology - WG**
  - C. Stoeckert
  - U.Penn
  - Standard for experiment annotation

- **Normalization - WG**
  - J. Quackenbush
  - TIGR
  - Standard for normalization methods

- **ToxicoGn - WG**
  - S.Sansone
  - EBI
  - Standard for contextual information
Standard 1: What to store?

The MIAME Requirements
Defining the Minimum Information About a Microarray Experiment

  - Defining the critical domains of a microarray experiment requiring sufficient annotation to be provided
  - Set of guidelines issued for the microarray community
  - Nailing down the rationales for accurate recording
  - Providing a framework to start from for establishing public repositories
  - Insisting on a need for infrastructure for data sharing
What does MIAME say?

- **Sufficient information** must be recorded to:
  - Correctly interpret and verify the results
  - Replicate the experiments

- **Structured information** must be recorded to:
  - Query and correctly retrieve the data
  - Analyse the data
6 parts of a microarray experiment

- Sample source
- Sample treatments
- Metadata
- Extraction protocol
- Labeling protocol

MIAME principles

- Hybridisation protocol
- Array design (Location and description of each element)
- Quantification matrix
- Analysis protocol
- Software specifications

- Sample
- Hybridisation
- Array

- Image
  - Scanning protocol
  - Software specifications

(S. Sansone)
6 parts of a microarray experiment

MIAME principles

- Type
- Factor
- Quality control
- Publication

- Strategy
- Algorithm
- Array control elements

- Data processing levels
- Lack of gene expression measurement units!
Standard 2: How to efficiently annotate data?

=> The MGED ontology (MO)

- MIAME publication evokes the need for efficient annotation
  - Reducing use of synonyms or ambiguous terms
  - Avoiding free text description
  - Use of annotation good practice and use of structured controlled vocabularies and available ontologies

- Definition: An Ontology is a ...
  - Domain specific dictionary capturing semantic relationship between terms
Standard 2: How to efficiently annotate data?

=> The MGED ontology (MO)

- Effort coordinated by C. Stoeckert & H. Parkinson
  - Placed under the GOBO umbrella, a stable version has now been released
  - MO is registered as GO Xref and at MOBY
  - Opensource, modular and interconnected to existing ontologies
    - Avoiding competing or redundant work
    - Join and participate basis
  - A set of rules on how to use it have been presented during MGED6 meeting

MGED ontology available from: http://mged.sourceforge.net/
Standard 3: How to store microarray data?

=> The MAGE-OM

- Second Milestone delivered by the MGED society
- Effort coordinated by Paul Spellman
- Model developed by Ugis Sarkans at the EBI
- Joined submission by MGED / Rosetta (Michael Miller)
- Officially approved by the OMG in October 2002
- Model now « frozen » for 2 years
MAGE-OM: an overview

- MAGE-OM a formalized representation of the world of microarray independent of
  - Experimental platform
  - Image analysis method
  - Normalization method

- MAGE-OM contains 16 main groups (or packages)

  - In the UML formalization:
    - Objects are described by Classes
    - Related classes are grouped together into packages
MAGE Object Model made simple

- **What was used**
  - BioMaterial
  - Array
  - ArrayDesign
  - BioSequence

- **What was done**
  - Experiment
  - BioAssay

- **Results**
  - HigherLevelAnalysis
  - BioAssayData
  - QuantitationType

- **Miscellaneous**
  - Measurement
  - AuditAndSecurity
  - Description
  - Protocol
  - BioEvent
  - BQS

(U. Sarkans)
MAGE Object Model: zooming in...
Standard 4: How to exchange microarray data?

=> The MAGE-ML format

- What is MAGE-ML? A byproduct of the MAGE-OM
  - Basically, it’s an XML file:
  - The corresponding DTD has been automatically generated from the MAGE Object Model
  - it is (almost) human readable

- Why MAGE-ML?
  - Technically, a broad range of tools already available to handle, parse XML
  - Ease of use in object oriented programming software development environment
Oct 2002: Press release by Rosetta/Agilent:
  • Resolver 6.0 now accepts MAGE-ML format

NCI’s Director Challenge adopts MAGE-ML as internal data exchange format (Ken Buetow, SOFG meeting 2002)

BASE Lund University LIMS works on MAGE-ML export

NHIES & NCT: adoption of MAGE-ML for toxicogenomics data
Standard 5: How to make data comparable?

=> Normalization and data transformation

- Preliminary studies showed limitation in data comparison
- Need for an independent MGED Working Group
- Effort coordinated by J. Quakenbush


- Taking into account within experiment hybridization variability to allow for quantitation level comparison
- Defining the comparison procedures and algorithm
- Defining the necessary control element types, amount of replications
- Issuing recommendations on data comparison, data transformation and potentially on experimental design
MGED standards -Summary

- Standard for **quantitative information**
  - MIAME

- Standard for **qualitative annotation**
  - MGED Ontology

- Standard for **data representation/exchange**
  - MAGE-OM & derived MAGE-ML language

- Standard for recording **controls, normalization** methods
December 2002: A New Publication from the MGED society

- Nature and related Journals of the Nature Publishing Group
- Bioinformatics
- Lancet
- Science

Guide to authors, reviewers and editors of microarray gene expression papers.

More journals ask for accession number as prerequisite for submission!
In the (MGED) standards we trust...

MGED standards
1-MIAME
2-MAGE-OM
3-MAGE-ML
4-MO

Regulatory networks
Tumor classifier
Standardization efforts
  • Microarray standards by the MGED Society

Infrastructure at the EBI
  • ArrayExpress Functional implementation of MGED standards
  • ExpressionProfiler

Future and Development
  • Infrastructure
  • standards
Infrastructure @ EBI

DATA SUBMISSION

MIAMExpress (MySQL)

Repository

ArrayExpress (Oracle)

MAGE-ML

Data matrix

Expression Profiler

MAGE-ML

Other Bioinformatic Databases @ EBI

DATA QUERY

DATA ANALYSIS

www

www

www

www

www

www

www

DATA IMPORT/EXPORT

DATA SUBMISSION

MIAME-compliant MAGE-ML pipelines
(Affymetrix, Agilent, NIEHS, NCI, TIGR, SMD, Sanger Institute, Base, Maxload)

MIAMExpress (Oracle)

Expression Profiler

Data Analysis Softwares (BioConductor J-express)

Other Microarray and Toxicogenomics Database
ArrayExpress: Public Repository Infrastructure

- MAGE-OM Implementation in Oracle 9i
- 132 SQL tables: oracle/SQL scripts available at:
  www.ebi.ac.uk/microarray/ArrayExpress/Implementation/implementation.html

- DB tools: (available from the same URL)
  - Validator: checking datafiles against our implementation of the OM
  - Loader/Unloader: input is a MAGE-ML file
  - Query interface: data access and retrieval
    (Java servlets on Tomcat server)
Submission to ArrayExpress (part 1)

- Direct submission as MAGE-ML documents:

  Creation of pipelines from local /private databases/LIMS:

  - Commercial resources:
    - Affymetrix GDAC-Exporter SDK.
    - Oracle 9i XDK
  
  - Open source resources:
    - MAGE-Stk Api (Perl and Java)

  EBI MAGE Validator/Loader distributed

  Valid MAGE-ML file
Submission to ArrayExpress (part 2)

Web based submission using MIAMExpress:

- **URL**: [www.ebi.ac.uk/miamexpress/](http://www.ebi.ac.uk/miamexpress/)
- Based on MIAME questionnaire
- Built-in Controlled Vocabulary
- Opensource Perl-CGI/mysql/Apache (GNU license)
- Usable as a Lab Notebook for daily data storage
- Automatically generates valid MAGE-ML documents
MIAMExpress Submission (part 1)

- MIAMExpress official release 15/12/2002:
  - Already 26 successful experiment submissions released
MIAMExpress Submission (part 2)

- User assistance and annotation standardization: EnsMART + ADF converter

### Genomic Annotation file

- **Nucleic Acid Component**
  - Ensembl Gene ids
  - Ensembl transcript ids
  - Embl-GenBank DDBJ ids
  - NCBI RefSeq ids
  - Locus Link ids

- **Protein Component**
  - Swall ids
  - Swissprot ids
  - PDB ids
  - Interpro ids
  - Ensembl Protein Family

- **Functional Component**
  - GO ids
  - GO descriptors + EvC

- **Phenotype Component**
  - OMIM id
ArrayExpress – Access to Data

- ArrayExpress Query interface:
  URL: www.ebi.ac.uk/arrayexpress/query/entry

- Providing private and public access to data: close cooperation with journals and editors
ArrayExpress DataSets

- In the Production Database: > 2700 hybridizations

Experiment Submissions broken down by Pipelines
- MIAMExpress 30%
- the Sanger Institute 16%
- RZPD Berlin 13%
- TIGR 13%
- CAGE 7%
- UMC Utrecht 7%
- NASC 7%
- Others 7%

Experiment Submissions Broken Down by Species
- Homo sapiens 17%
- Mus musculus 24%
- Rattus norvegicus 9%
- S. cerevisiae 6%
- S. pombe 6%
- Caenorhabditis elegans 6%
- Xylella fastidiosa 2%
- Apis mellifera 2%
- Drosophila melanogaster 2%
- Arabidopsis thaliana 19%
Talk structure

- Standardization efforts @ EBI
  - Microarray standards by MGED Society

- Standard-supporting microarray infrastructure@EBI

- Future and Development
  - Infrastructure
  - standards
New pipeline implementation with large centres  
  (e.g. MAGE-ML adopted by NCI for file exchange)

Development of Quality Control rating procedures  
  Creation of quality metrics to rate Experiment submission  
  Creation of MIAME validator

Development of appropriate curation tools  
  Automated submission tracking  
  Better integration with Ontology browsers  
  Improvement of the Validator to enhance CV checking
Online submission and MIAMExpress Mk II:

- *Scaled up to match the full MAGE model*
- *Integration of MGED ontology in the interface*
- *Extension according to the user needs & feedback*
- *Increase of usability, flexibility and scalability*
ArrayExpress’s Future: Output wise

- Development of graphical display interface
  - *Creation the interface for displaying results in a meaningful way.*

- Development of Gene-centric / Sequence-centric queries
  - *Several hurdles: Generating a “consensual” Gene Index*
  - *Standardizing Array Annotation: re-annotating Arrays?*
    - => unavoidable disclosure of sequences used in array
    - => IP issues
    - *Solution: follow Affymetrix example! working along with Ensembl!*

- Development of query storage/tracking tools
ArrayExpress’s Future: Output wise (2)

- a possible view of MAGE-ML based data exchange in action ...(if all repositories rely on standards)

**Array Express**

- Data Import from local databases/LIMS
- MAGE-ML
- Data Export to 3rd party tools

- Export to EBI local toolbox

- Japanese Repository Project
- Cibex
- Geo
- US Repository

- MAGE-ML

- US Repository

- MAGE-ML
ArrayExpress’s Future:

- Better integration with Expression Profiler
- Development of a Datawarehouse in a production scale (for the moment prototyping)

- New MAGE-OM (U. Sarkans, P. Spellman)
  - To ensure better description of Biological material used in experiments
  - Application to new emerging technologies in the field of high throughput biology (application to Proteomics)

- MIAME-Toxicogenomics guidelines:
  - Consequence of ILSI project and interest expressed by NHIES and FDA (sansone@ebi.ac.uk: Toxicogenomics coordinator @ EBI)
Microarray Informatics team at EBI

Alvis Brazma - group leader

ArrayExpress
• Gonzalo Garcia
• Ahmet Oezcimen
• Anjan Sharma
• Ugis Sarkans

MIAMExpress
• Mohammadreza Shojatalab
• Niran Abeygunawardena
• Sergio Contrino

Expression Profiler
• Misha Kapushesky
• Patrick Kemmeren
• Jaak Vilo

EU Temblor funding

Toxicogenomics / Nutrigenomics effort
• Susanna Sansone
• Philippe Rocca-Serra
• Sergio Contrino

Research
• Thomas Schlitt
• Katja Kivinen
• Lev Soinov
• Anastasia Samsonova
• Aurora Torrente

Curation
• Helen Parkinson
• Philippe Rocca-Serra
• Ele Holloway
• Gaurab Mukherjee
Reminders

- MIAME Checklist and standard enforcement policies
  - Journals
  - Funding agencies
- EBI: [www.ebi.ac.uk/microarray](http://www.ebi.ac.uk/microarray)
  - Data storage resources
    - MAGE-OM and MAGE-ML dtd
    - ArrayExpress: SQL scripts + Validator + MAGE-ML files
    - MIAMExpress: sourceforge project ID49908
- MGED: [www.mged.org](http://www.mged.org)
  - MGED Ontology GOBO project
  - Meetings:
    - MAGE jamboree meeting Hinxton 1-6 December 2003