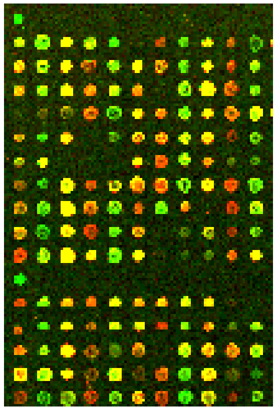
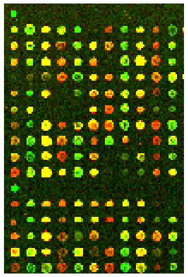
Data Quality in Microarray Databases

AMW07 – Octubre 2007
Punta del Este - Uruguay

Lorena Etcheverry
InCo - Facultad de Ingeniería
Universidad de la República

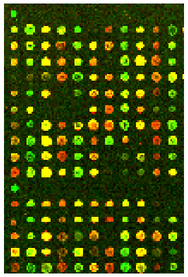


Agenda



- Motivation
- Biological Background
  - Basic definitions
  - Experimental data and storage formats
- Our proposal for quality evaluation
  - Definition of Quality factors and metrics
  - Composition tool
  - Examples

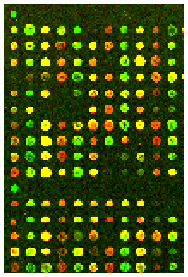
Motivation

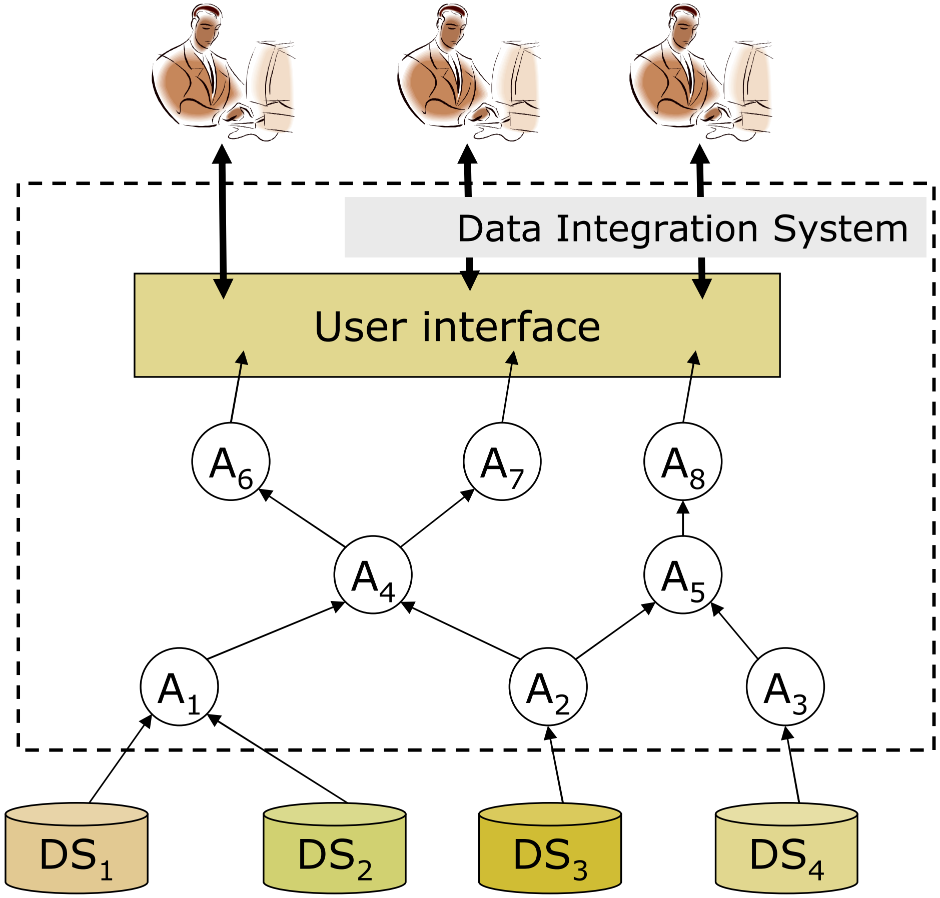


- “Biological research has transformed from a purely experimental to an information-driven discovery science” Markowitz, VLDB2004

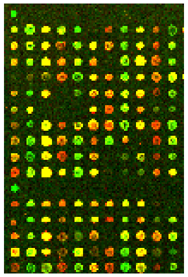
- “Data of poor quality in genomic databases have enormous medical and economical impact on their users/customers” Naumann, ICIQ03

Motivation



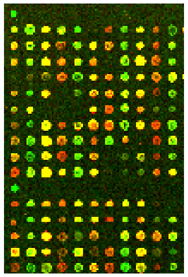

Data Integration System
User interface
A_6
A_7
A_8
A_4
A_5
A_1
A_2
A_3
DS_1
DS_2
DS_3
DS_4

Motivation

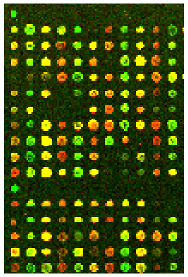


- Large data volumes
  - Lots of public experiments data (internet)
  - Average experiment size: 300 MB
  - Local data also (LIMS)
- Specific and complex formats
- Poor data quality
- No standard definitions of quality factors and metrics

Objectives



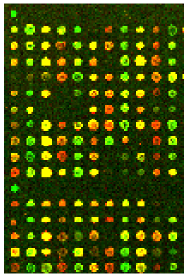
- Measure the quality of the experiments stored in the internet
- How?
  - Definition of quality properties
    - Specific to biological context
    - Adapted to concrete users
  - Developing a tool for aiding in the definition of quality properties



Agenda

- Biological Background
  - Basic definitions
  - Storage formats
- Our proposal for quality evaluation
  - Definition of Quality factors and metrics
  - Composition tool
  - Examples

Some basic concepts



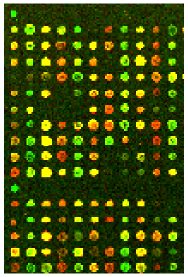
- **Genes and genome** (DNA):

  are the basic framework that determine the characteristics of individuals.

- **Proteins and proteome**:

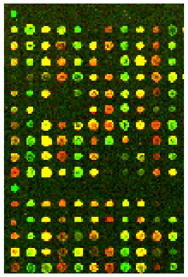
  in order to perform biological analysis, other information is also required, specially the protein structure of the individuals (proteome), which is generated from the genetic structure (genome).

Gene Expression



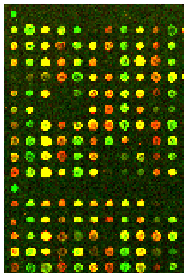
- Computing "**gene expression**" is a fundamental type of biological experiment, which goal is:
  - Determining "how active" is a gene (a sample).
- Application:
  - To discover genetic-based patterns of diseases:
    - which are associated to gene activity in cells (gene turned on/off)
    - which are identified by analyzing the expression of the genes in normal and in ill cells.

Example of gene expression application

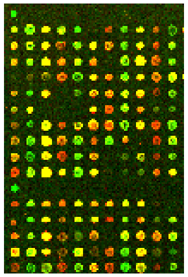


- *"Researchers compared the gene expression profiles of four different types of mouse skin cells: low-risk and high-risk papillomas (benign tumors) that had been chemically induced, as well as normal skin and cancerous cells. The investigators demonstrated that precancerous lesions can be separated into subgroups according to distinct patterns of gene activities — namely, which genes were turned on or off."*
- *"A specific pattern of activity, sometimes called a molecular signature, was present in the precancerous lesions, and correlated with a higher risk for malignant conversion"*. Oncogene (21 May 2007)

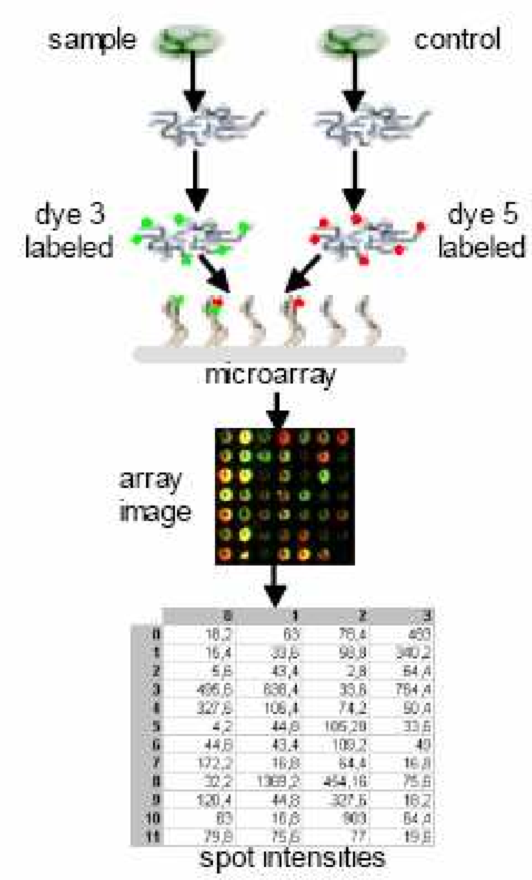
Gene Expression and Microarrays



- What is a microarray?
  - a collection of microscopic DNA spots arrayed as a matrix on a solid surface.
  - Over 60.000 spots in each array.
- What are they used for?
  - allows large-scale **gene expression** study and comparison
  - high-throughput technique (lots of data)



How does it work?



	0	1	2	3
0	18,2	63	78,4	483
1	16,4	33,6	98,8	340,2
2	5,6	43,4	2,8	64,4
3	495,6	638,4	33,6	784,4
4	327,6	106,4	74,2	50,4
5	4,2	44,8	105,28	33,6
6	44,8	43,4	109,2	40
7	172,2	16,8	64,4	16,8
8	32,2	1369,2	454,16	75,6
9	120,4	44,8	327,6	18,2
10	63	16,8	903	64,4
11	79,8	75,6	77	19,6

(1) Cell selection

↓

(2) RNA/DNA preparation

↓

(3) Hybridization

↓

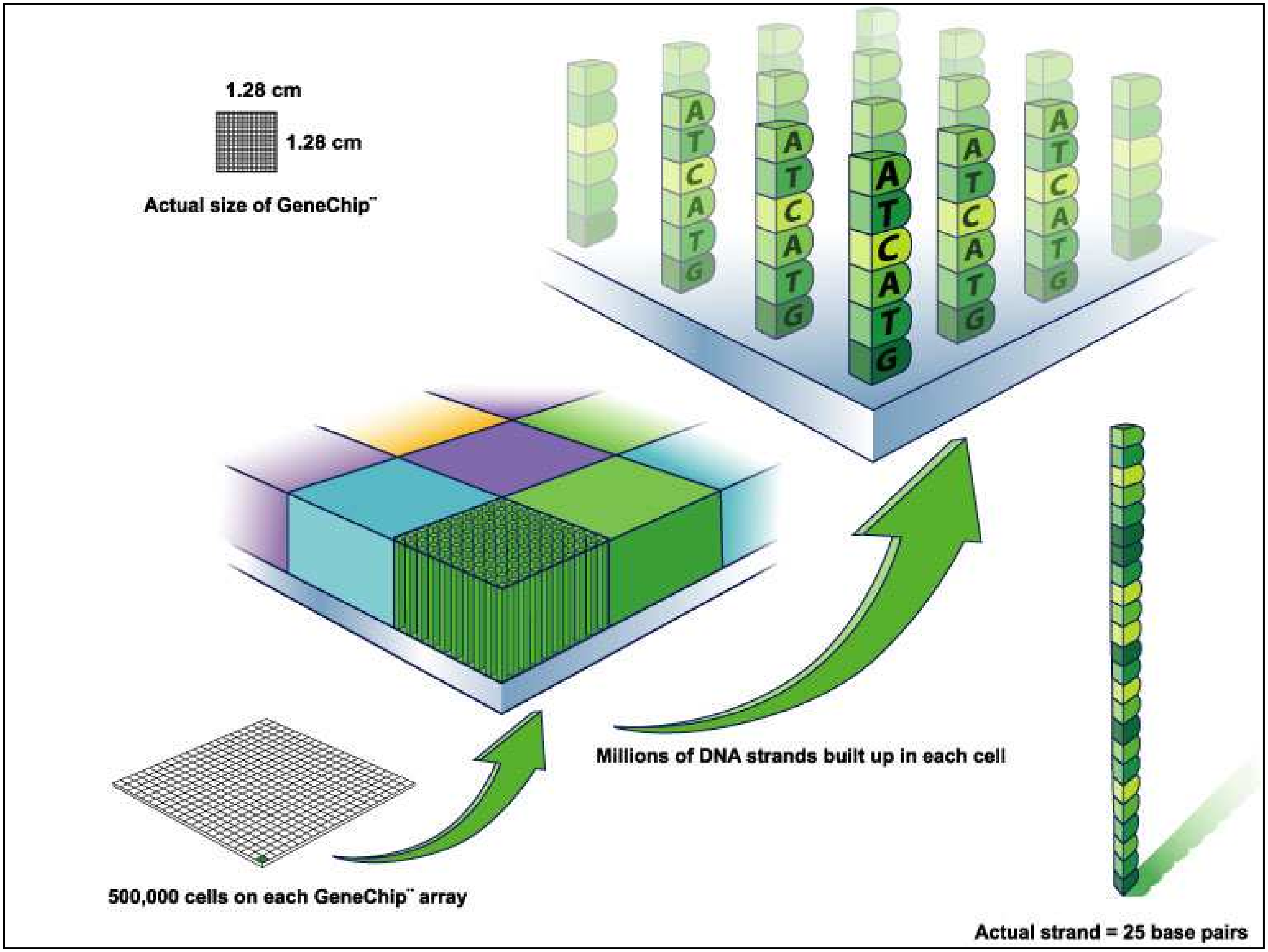
(4) Array scan

↓

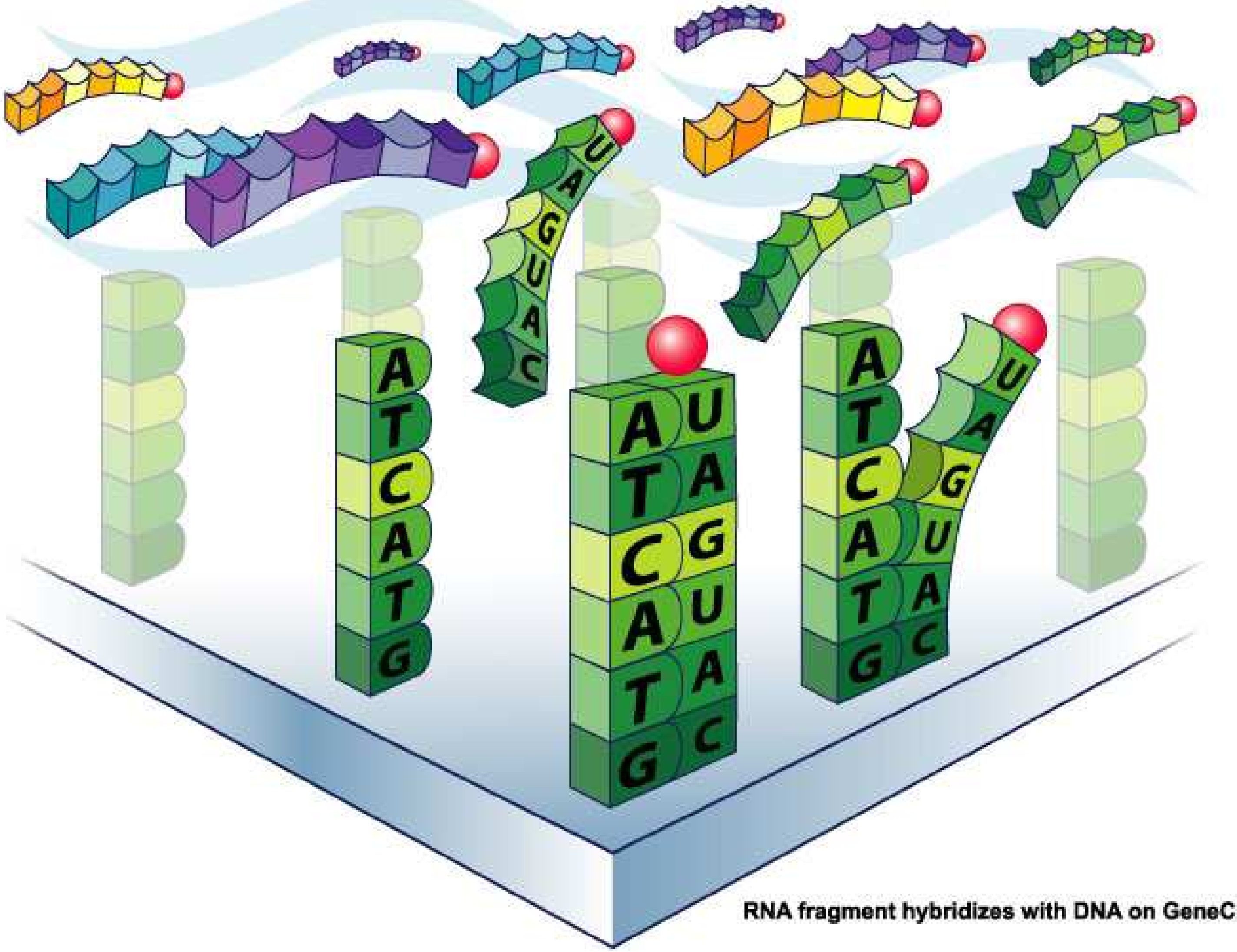
(5) Image analysis

↓

(6) Expression analysis

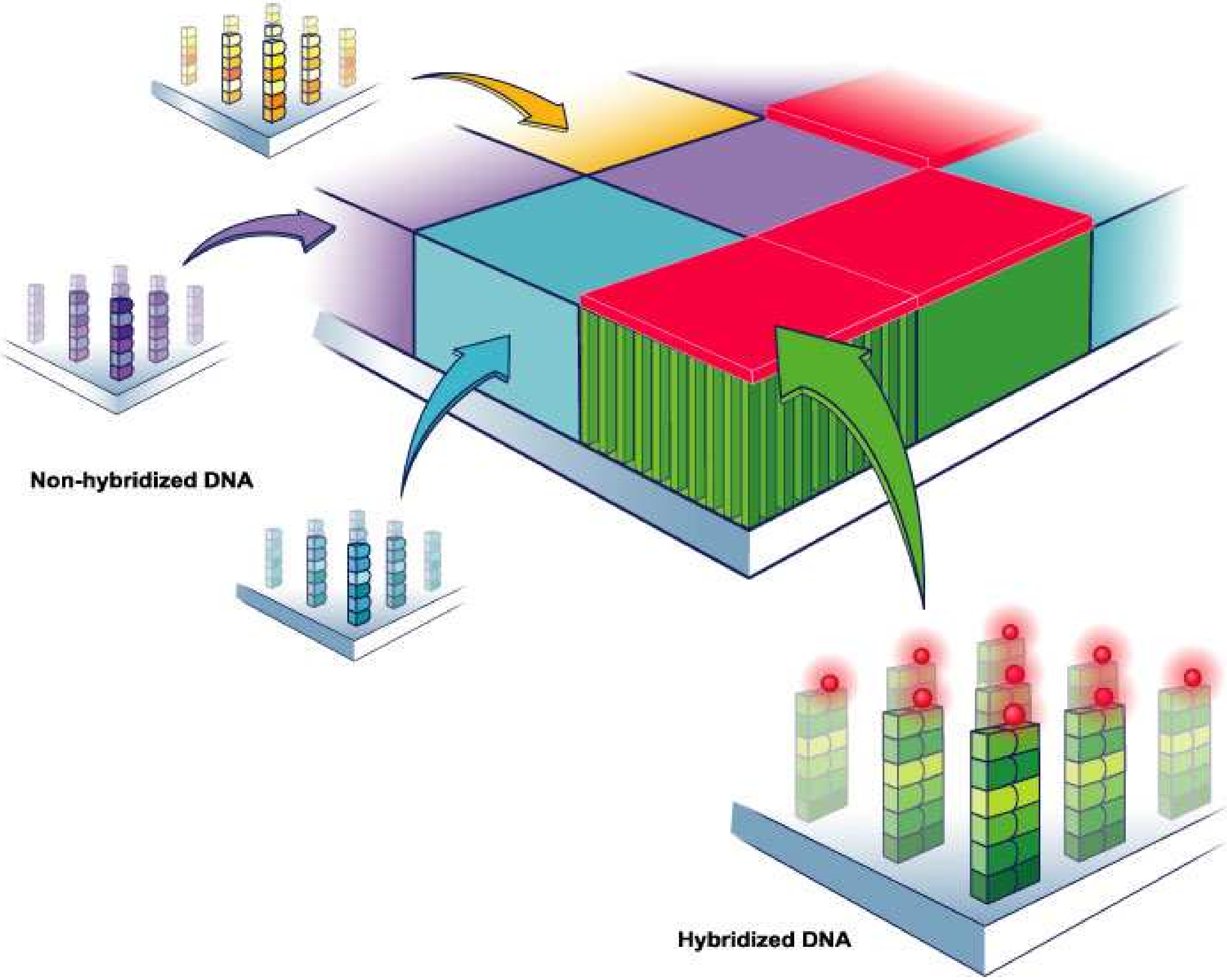

1.28 cm
1.28 cm
Actual size of GeneChip™
A
T
C
A
T
G
500,000 cells on each GeneChip™ array
Millions of DNA strands built up in each cell
Actual strand = 25 base pairs

RNA fragments with fluorescent tags from sample to be tested



RNA fragment hybridizes with DNA on GeneChip

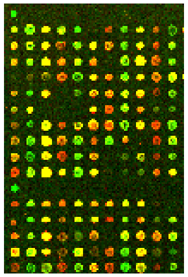
Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow

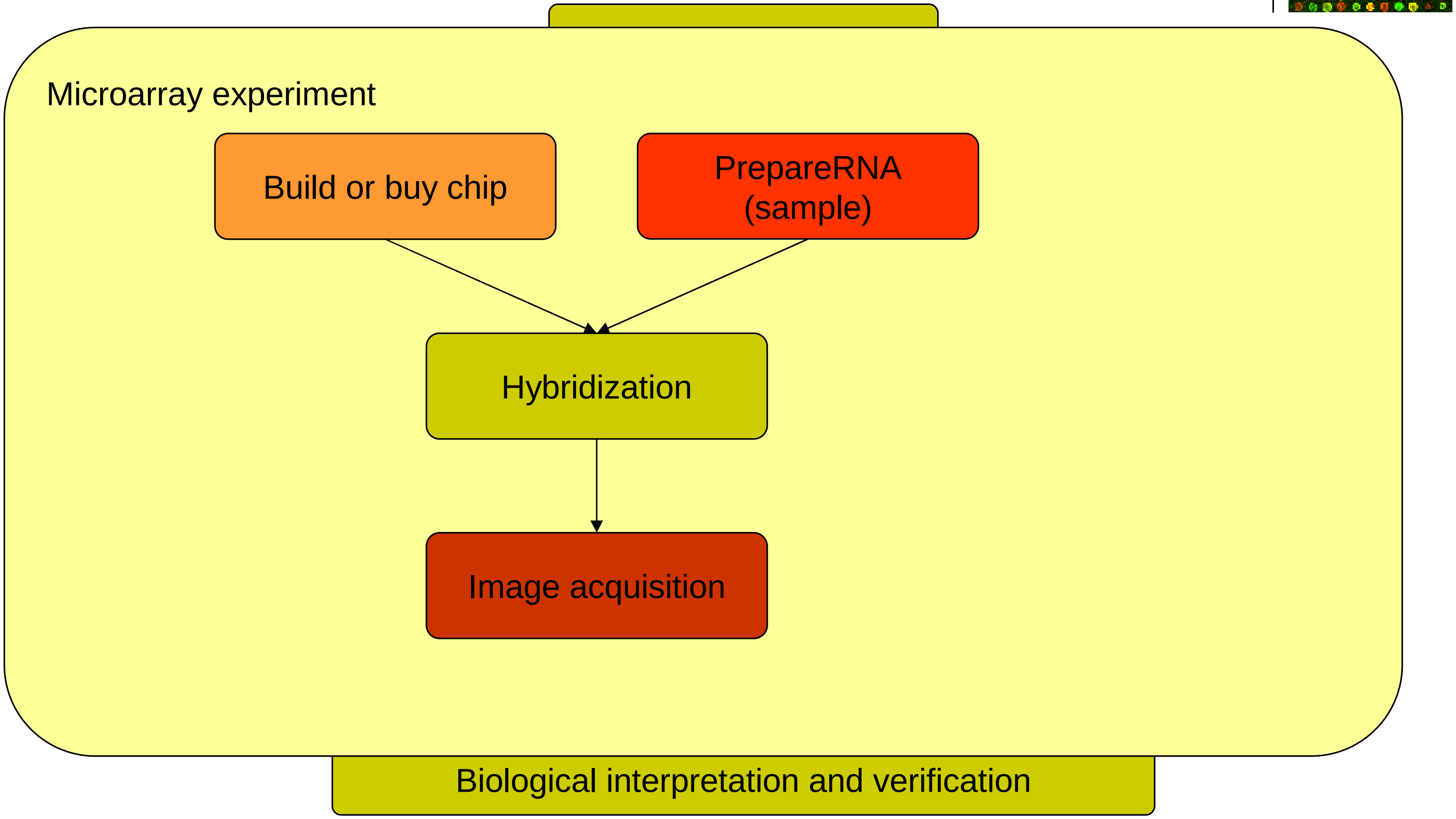


Non-hybridized DNA

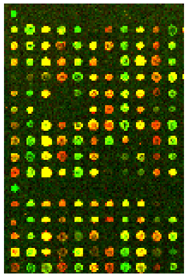
Hybridized DNA

Microarray experiments are processes!





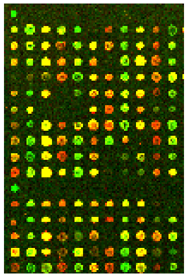
Agenda



- Biological Background
  - Basic definitions
  - Experimental data and storage formats

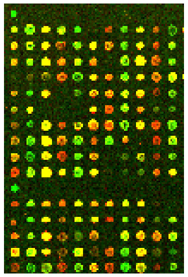
- Our proposal for quality evaluation
  - Definition of Quality factors and metrics
  - Composition tool
  - Examples

Dealing with microarray experimental data



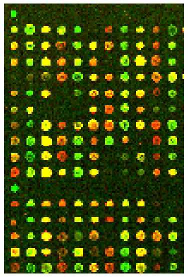
- Stored data provides detailed information about gene expression experiments.
- Kinds of data (resulting from the different steps):
  - Raw images.
  - Raw numerical data : spots intensity
  - Preprocessed numerical data.
    - Raw numerical data - noise + statistical processing (ex. normalization)
- How is this data used?:
  - Users take the preprocessed numerical data and perform statistical analysis in order to proof their hypothesis (hypothesis tests)

Sources of microarray experimental data



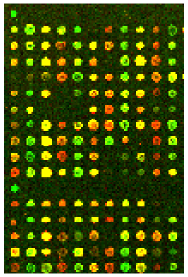
- There are over 10 different public repositories [Lemoine02], [Do03].
- “to encourage and empower biologists to provide results in a structured and computable format alongside publication” [Boguski01]
- The MGED group suggests that **journals require submission** of microarray data to either of two databases emerging as the main public repositories: GEO or ArrayExpress.

Microarray experiments data standards

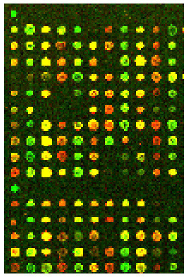


- MAGE-OM:
  - object model written in UML
- MAGE-ML:
  - exchange format for MAGE-OM
  - Implemented as an XML dtd
- MGED Ontology:
  - Controlled vocabulary
  - Implemented in OWL

MAGE-OM



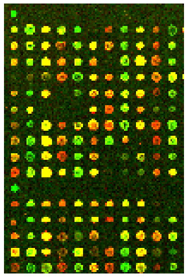
- Allows the representation of:
  - Experiment description (experiment metadata)
  - Experiment results
- Complex data-centric model:
  - 132 classes
  - 17 packages
  - 223 associations between classes.
- Accepted by the OMG as a biosciences standard.
- Each repository uses it's own relational implementation of the model. DRAWBACK



MAGE-ML

- Adopted as exchange format by all major public repositories.
- We have decided to base our proposal in this format in order to be able to **integrate data** from different repositories

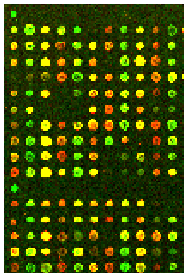
MGED Ontology



- An ontology for microarray experiments description
  - particular emphasis on biological material (biomaterial) annotation.
- Purpose: provide standard terms for the annotation of microarray experiments [Whetzel06]

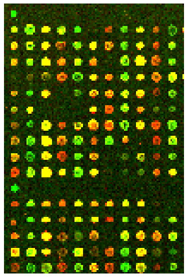
- Provides a controlled vocabulary
- Presents many **design problems** [Soldatova05]
  - Can’t be used as a classifier
  - Can't be used to infer.

How should microarray experimental data be stored?



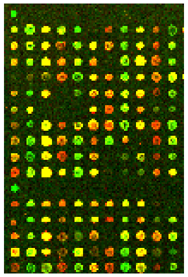
- Experiment repositories should support this standards.
- Only 3 of them really implement standards [Do03]:
  - ArrayExpress (European Bioinformatics Institute)
  - GEO (National Center for Biotechnology Information, US)
  - SMD (Stanford University, US)

Agenda



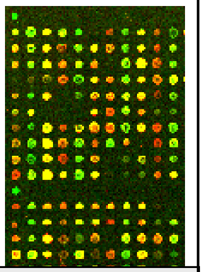
- Biological Background
  - Basic definitions
  - Experimental data and storage formats
- Our proposal for quality evaluation
  - Definition of Quality factors and metrics
  - Composition tool
  - Examples

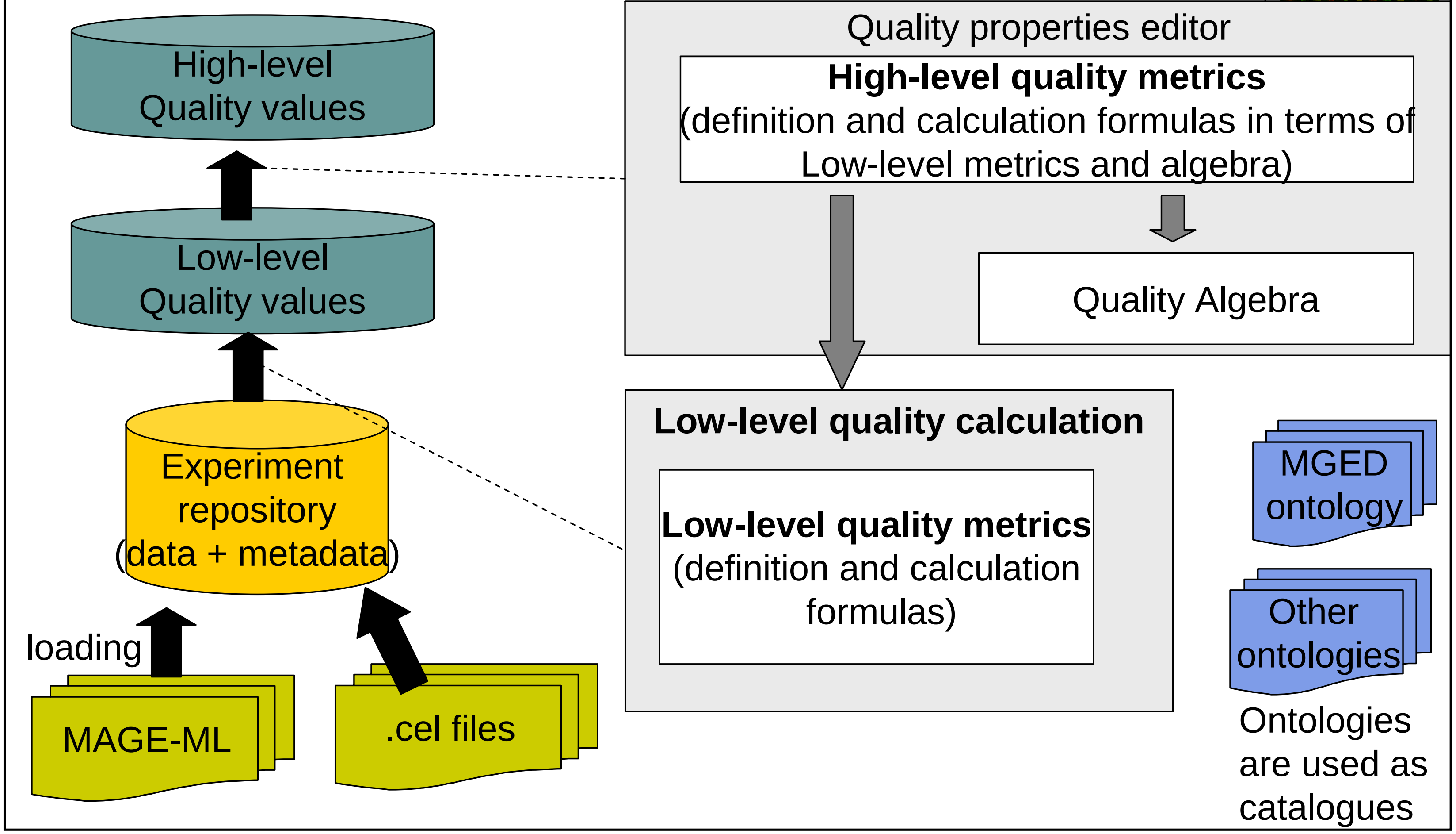
Measuring quality of a microarray experiment



- Our approach:
  - Identify “low-level” quality metrics.
    - Experiment quality metrics (biological point of view)
    - Experiment description quality metrics (“data quality” point of view)
  - Measure quality according to these metrics.
  - Allow the end-user to build his own “high-level” quality metrics, based on:
    - Low-level metrics
    - Some kind of “quality algebra” (composition, aggregation)

Proposed architecture





High-level Quality values

Low-level Quality values

Experiment repository (data + metadata)

loading

MAGE-ML

.cel files

Quality properties editor

High-level quality metrics
(definition and calculation formulas in terms of Low-level metrics and algebra)

Quality Algebra

Low-level quality calculation

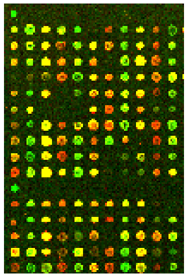
Low-level quality metrics
(definition and calculation formulas)

MGED ontology

Other ontologies

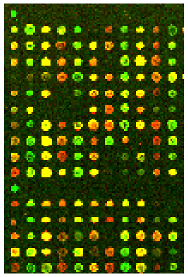
Ontologies are used as catalogues

Experiment quality metrics



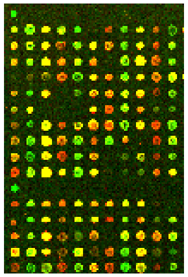
- According to biological literature there are several aspects that affect experiment quality.
- This aspects, if present in the model, could be considered as the “building blocks” of high level quality properties.

Experiment quality metrics (2)



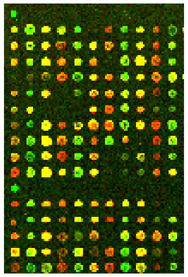
- Biological sample quality
- Experimental design quality
- Preprocessed data quality (numerical results)
- Data interpretation quality

Experiment quality metrics (3)



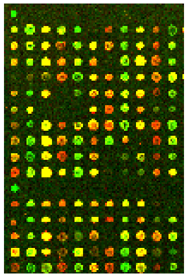
- Biological sample quality:
  - Sample quality can be quantitatively assessed.
    - RIN number
    - RNA degradation plot (gradient)
- Experimental design quality
  - What type of experiment is being performed? (dose response,
  - Replication type (biological/technical)
  - Is pooling being done?
  - How many individuals per biological replicate?
  - How long did they take to make the experiment?
  - How many different operators have been involved?

Experiment quality metrics (4)

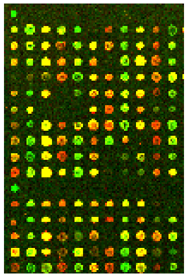


- Preprocessed data quality:
  - What algorithm has been used to subtract image noise?
  - What normalization method?
- Interpretation quality:
  - What statistic method are they using to inference the results?
    - Fold-change is not good enough.
    - Variance shrinkage methods lead to better results.

Experiment Description Quality Metrics



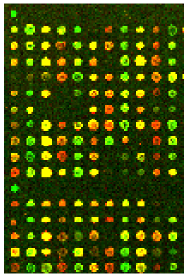
- Accuracy
  - Syntactic correctness
  - Precision
- Consistency
- Completeness
- Freshness



Syntactic correctness

- Experiments contains references to individuals in:
  - MGED Ontology
  - Other ontologies (NCBI taxonomy, etc.)
- In MAGE-ML they are tagged as <*OntologyEntry*>
- Every ontology entry should reference an individual in that ontology.

Syntactic correctness (2)



- Example:

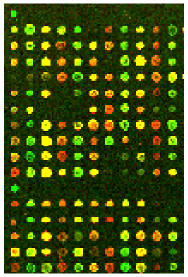
```
<QualityControlDescription_a
 <Description>
  <Annotations_assnlist>
   <OntologyEntry
    value="Biological replication”
  category="QualityControlDescriptionTyp
  e">
   </OntologyEntry>
   </Annotations_assnlist>
  </Description>
</QualityControlDescription_assn>
```

“value” should be an individual of the class QualityControl DescriptionType

WRONG

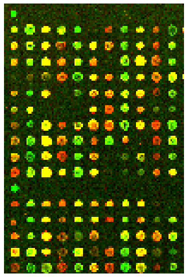
MGED ONTOLOGY
biological_replicate
technical_replicate
peer_review_quality_control
spike_quality_control
dye_swap_quality_control
real_time_PCR_quality_control
reverse_transcription_PCR_quality_control

Precision



- Hierarchical data: level in the hierarchy.
- Example:
  - Description of the source of the biological sample.
    - Organism, often is a reference to an external taxonomy.
    - Mouse, mouse kidney, mouse kidney epithelial cell

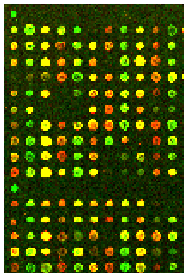
Consistency

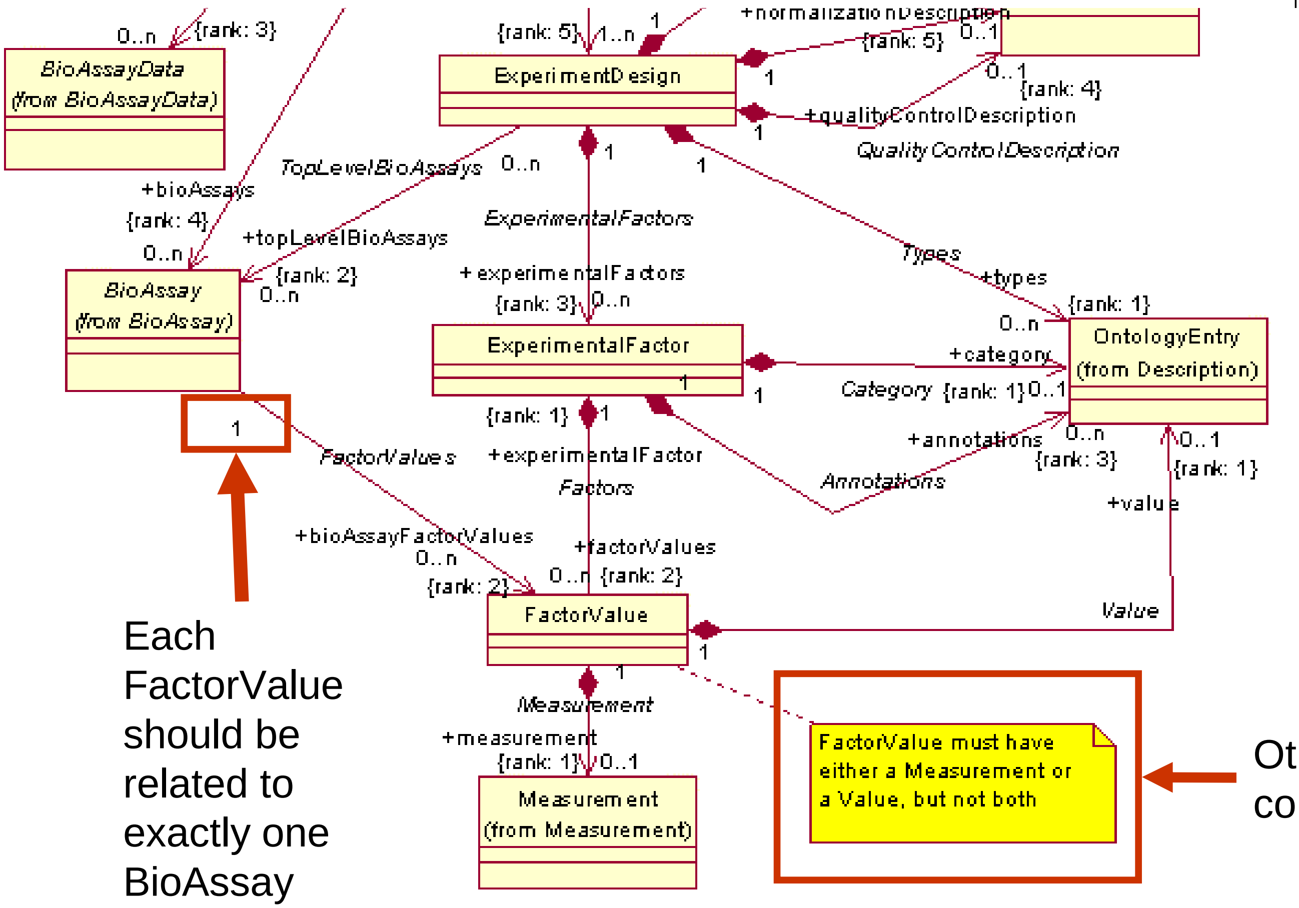


- MAGE-OM constraints (∃, ∀, cardinality) should be tested over MAGE-ML documents or relational data.
- “Well formed” experiments

- The MAGE-OM model also contains “free text” constraints that should be checked (not even OCL)

Consistency (2)



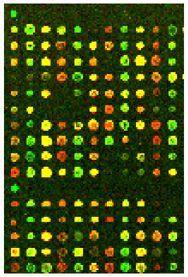


FactorValue must have either a Measurement or a Value, but not both

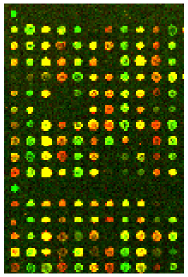
Each FactorValue should be related to exactly one BioAssay

Other constraints

Completeness



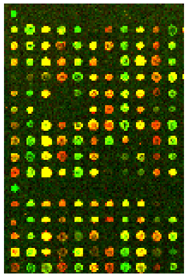
- The idea of how complete is the description of the experiment is very important in this context.
- Completeness could be defined in terms of how many parts of the model are present (and not empty) in the description.



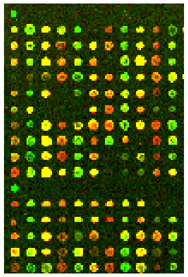
Freshness

- Experiment realization date
- Experiment submission date

High Level Quality Metrics



- End-user should be able to define his own quality metrics, based on his preferences and experience
- The value of the high level quality metrics should be calculated based on the formula that defines it.



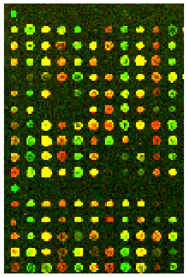
High Level Quality Metrics (2)

- Some examples

  -
    ```
    Quality = If (experiment_type =
    “dose_response” and biological_replication
    =“yes” and syntactic_corr > 75%)
          then quality=1
          else quality=0
    ```

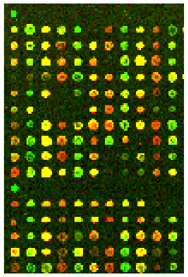
  -
    ```
    Quality = 1 – (syntactic_corr *
    completeness)
    ```

  -
    ```
    Quality = If (laboratory IN trusted_labs)
          then quality =1
          else quality =0
    ```



Current and future work

- Survey of low-level quality metrics
- Acquisition of source data (local experiment repository)
- Acquisition of low-level metrics
- A tool for aiding in the definition of high-level quality metrics
- Calculation of high-level metrics
- Begin to explore relations between quality factors.



References

- *[Allison06]* Allison et al. **“ Microarray data analysis: from disarray to consolidation and consensus “**. In: Nature Reviews Genetics, 7, 55-65 (January 2006)
- *[Boguski01]* Boguski, M et al **“Scientists discuss ongoing efforts to standardize and compare microarray expression data ”** BioOnline 2001 ,
- *[Do03]* Do, H.H. et al. **“Comparative Evaluation of Microarray-based Gene Expression Databases”,** Proc. 10. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW 2003)
- *[Lemoine02]* Lemoine, **“ Une évaluation globale des bases de données dédiées aux puces à AND** “ [on-line] , Service de Génomique Fonctionnelle CEA/Genopole d'Evry (2002)
- *[Nature02]* **“Microarray standards at last”** Nature, 2002. 419(6905): p. 323.
- *[Naumann03]* Naumann,F. et al. **“Data quality in genome databases”**;Proceedings of 8th International Conference on Information Quality ICIQ,2003
- *[Soldatova05]* Soldatova, L. et al. **“Are the current ontologies in biology good ontologies?”**, Nature Biotechnology 23, 1095 - 1098 (2005).