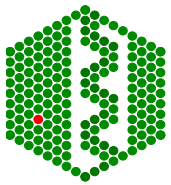
ArrayExpress – a public database for microarray gene expression data

Helen Parkinson

Microarray Informatics Team

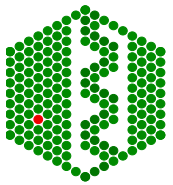
European Bioinformatics Institute

Transcriptome 2002, Seattle



Why have a public database?

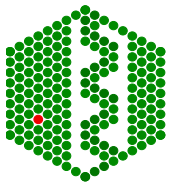
- Easy data access
- Resolves local storage issues
- Common data exchange formats can be developed
- Improved data comparison
- Curation can be applied
- Annotation can be controlled
- So that a public standard can be applied (peer review) – MIAME
- Additional info can be stored that is missing in publications



Or, to put it another way

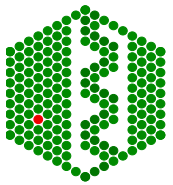
"…to encourage and empower biologists to provide results in a structured and computable format alongside publication"

Mark Boguski



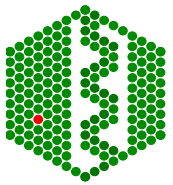
Talk structure

- MIAME standard
- Sample description and annotation Ontologies
- ArrayExpress
- Submission and annotation tool
- The future



Problems of microarray data analysis

- Size of the datasets
- Different platforms - nylon, glass
- Different technologies on platforms- oligo/spotted
- Referencing external databases which are not stable
- Sample annotation
- Array annotation
- Need for LIMS systems and the need for bioinformaticians

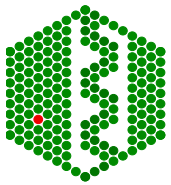


Standardisation of microarray data and annotations -MGED group

The goal of the group is to facilitate the adoption of standards for DNA-array experiment annotation and data representation, as well as the introduction of standard experimental controls and data normalisation methods.

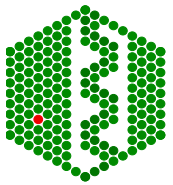
www.mged.org

Includes most of the worlds largest microarray laboratories and companies (TIGR,Affymetrix,Stanford,Sanger,Agilent, Rosetta, etc)



Glossary

- MIAME is a standard
- MAGE-OM is an object model
- ArrayExpress is a database implementation which uses that model
- MAGE-ML is a mark-up language auto generated from MAGE-OM
- MIAMExpress is a tool for generating data in MAGE-ML format

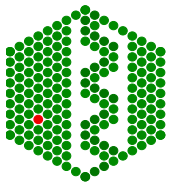


General MIAME principles

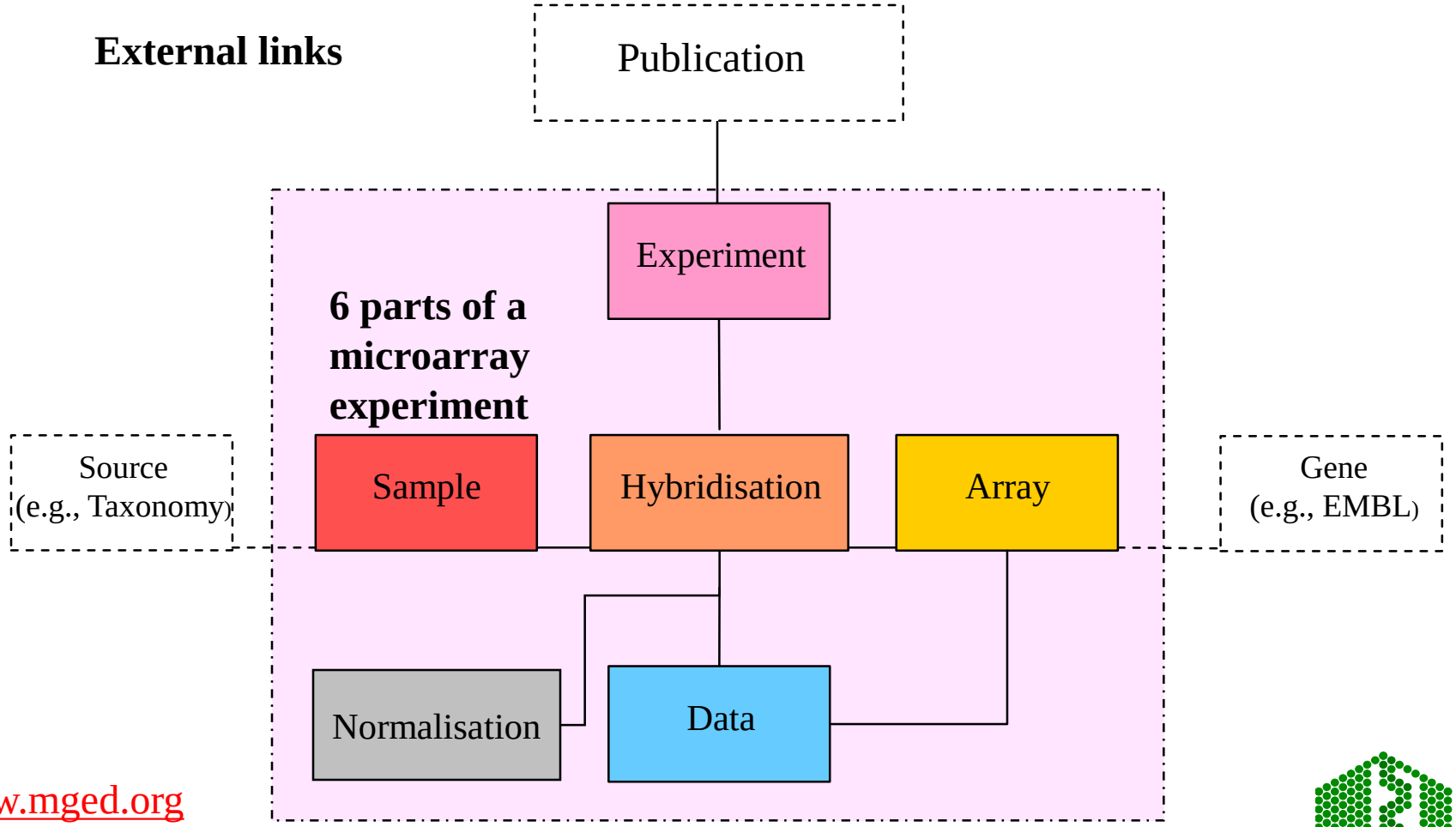
- Recorded info should be sufficient to interpret and replicate the experiment

- Information should be structured so that querying and automated data analysis and mining are feasible

Brazma *et al*,.Nature Genetics, 2001



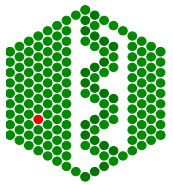
MIAME – Minimum Information About a Microarray Experiment



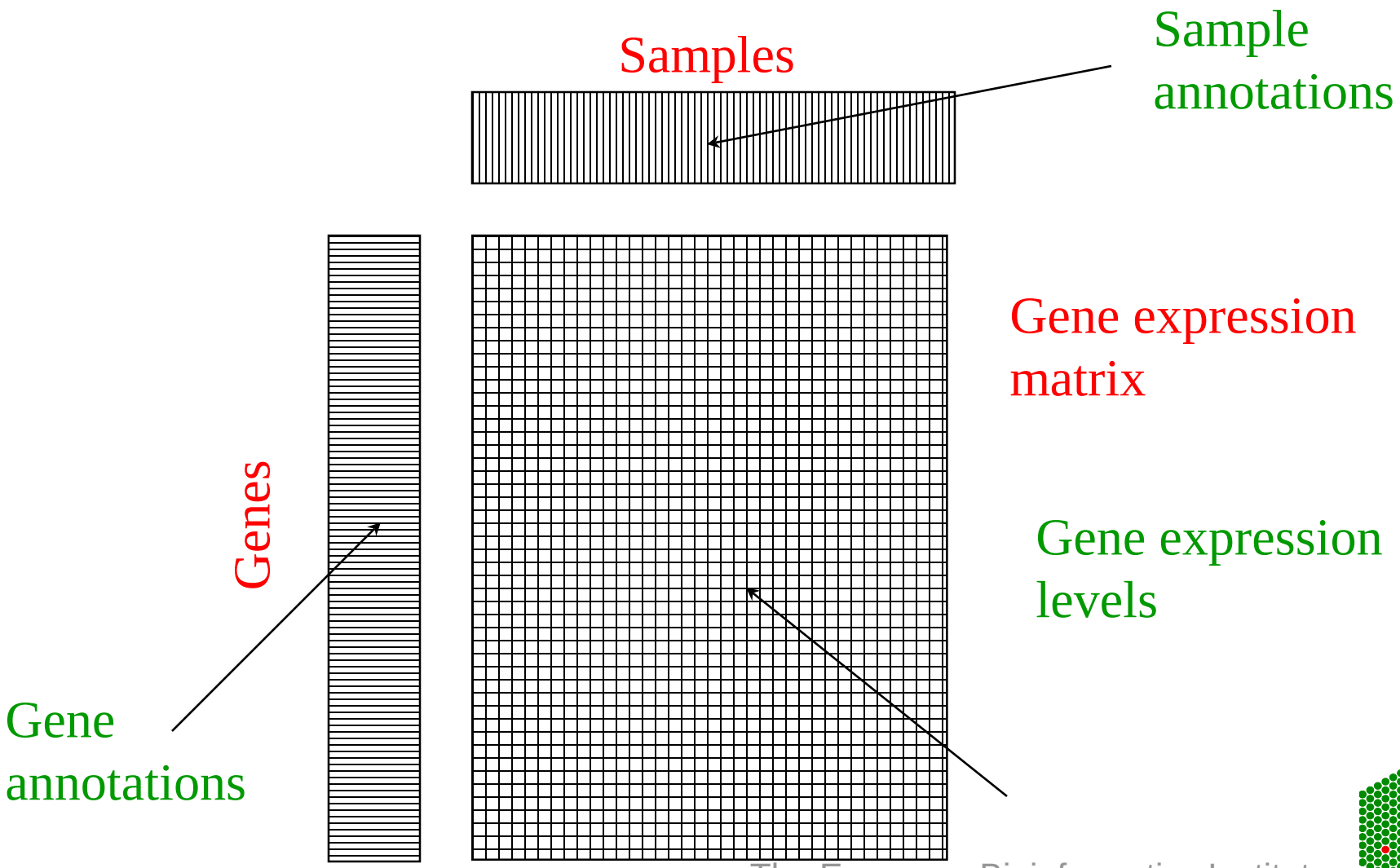
www.mged.org

The annotation challenge

- Use of controlled terms
- Data curation at source (LIMS)
- Avoidance of free text
- Integration of terms into query interfaces
- Removal of synonyms/or use of synonym mappings
- Provision of definitions and sources

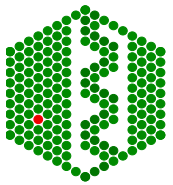


A gene expression database from the data analyst's point of view



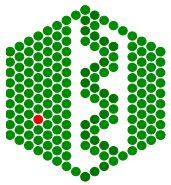
Gene Annotation

- Can be given by links to gene sequence databases and GO can be used on the analysis side
- MIAME is flexible, allows many kinds of sequence identifiers or even sequence itself
- In some cases it’s more useful to include a real sequence than an inaccurate id
- Submitters are encourage to submit seqs to public databases



Sample annotation

- Gene expression data only have meaning in the context of detailed sample descriptions
- If the data is going to be interpreted by independent parties, sample information has to be searchable and in the database
- Controlled vocabularies and ontologies (species, cell types, compound nomenclature, treatments, etc) are needed for unambiguous sample description
- These resources need mapping



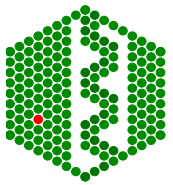
What Does an Ontology Do?

- Captures knowledge
- Creates a shared understanding – between humans and for computers
- Makes knowledge machine processable
- Makes meaning explicit – by definition and context
- It is more than a controlled vocabulary

From Building and Using Ontologies, Robert Stevens, U. of Manchester

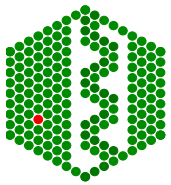
Examples of usable external ontologies

- NCBI taxonomy database
- Jackson Lab mouse strains and genes
- Edinburgh mouse atlas anatomy
- HUGO nomenclature for Human genes
- Chemical and compound Ontologies
- TAIR
- Flybase
- GO (www.geneontology.org)



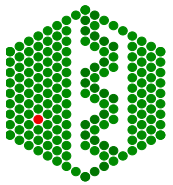
Sample annotation- what can be done?

- Build an ontology for gene expression data (MGED)
- Incorporate the ontology into the database and tools
- Use existing ontologies
- Develop internal editing tools for the ontology
- Develop browser or other interface for the ontology and link to LIMS
- Some use of free text descriptions are unavoidable (curation workload)



MGED Biomaterial (sample) Ontology

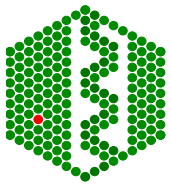
- Under construction – by MGED ontologists
  - Using OILed (though other tools exist)
- Motivated by MIAME and coordinated with the database model (mapping available)
- We are extending classes, provide constraints, define terms, provide new terms and develop cv's for submissions
- Other ontologies are under development, MAGE-OM ref's ontologies in ~50 places, these are being added to the MGED effort



Excerpts from a Sample Description

courtesy of M. Hoffman, Lion BioSciences

Organism: Mus musculus [NCBI taxonomy browser]
Cell source: in-house bred mice (contact: person@somewhere.ac.uk)
Sex: female [MGED]
Age: 3 - 4 weeks after birth [**MGED**]
Growth conditions: normal
controlled environment
20 - 22 ºC average temperature
housed in cages according to EU legislation
specified pathogen free conditions (SPF)
14 hours light cycle
10 hours dark cycle
[Developmental stage]: stage 28 (juvenile (young) mice)) [**GXD "Mouse Anatomical Dictionary"]**
Organism part: thymus [**GXD "Mouse Anatomical Dictionary"**]
Strain or line: C57BL/6 [**International Committee on Standardized Genetic Nomenclature for Mice**]
Genetic Variation: Inbr (J) 150. Origin: substrains 6 and 10 were separated prior to 1937. This substrain is now probably the most widely used of all inbred strains. Substrain 6 and 10 differ at the H9, Igh2 and Lv loci. Maint. by J,N, Ola. [**International Committee on Standardized Genetic Nomenclature for Mice**]
Treatment: in vivo [**MGED**] [**intraperitoneal**] **injection of** [**Dexamethasone**] **into mice, 10 microgram per 25 g bodyweight of the mouse**
Compound: drug [**MGED**] **synthetic** [**glucocorticoid**] [**dexamethasone**]**, dissolved in PBS**



Part of the MGED biomaterial ontology

class Age

documentation:

The time period elapsed since an identifiable point in the life cycle of an organism. If a developmental stage is specified, the identifiable point would be the beginning of that stage. Otherwise the identifiable point must be specified such as planting.

type:

primitive

superclasses:

BiosourceProperty

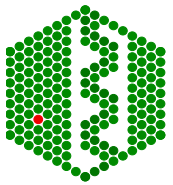
constraints:

slot-constraint has_measurement has-value Measurementslot-constraint initial_time_point has-value one-of (planting

beginning_of_stage)

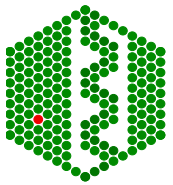
used in slots:

initial_time_point

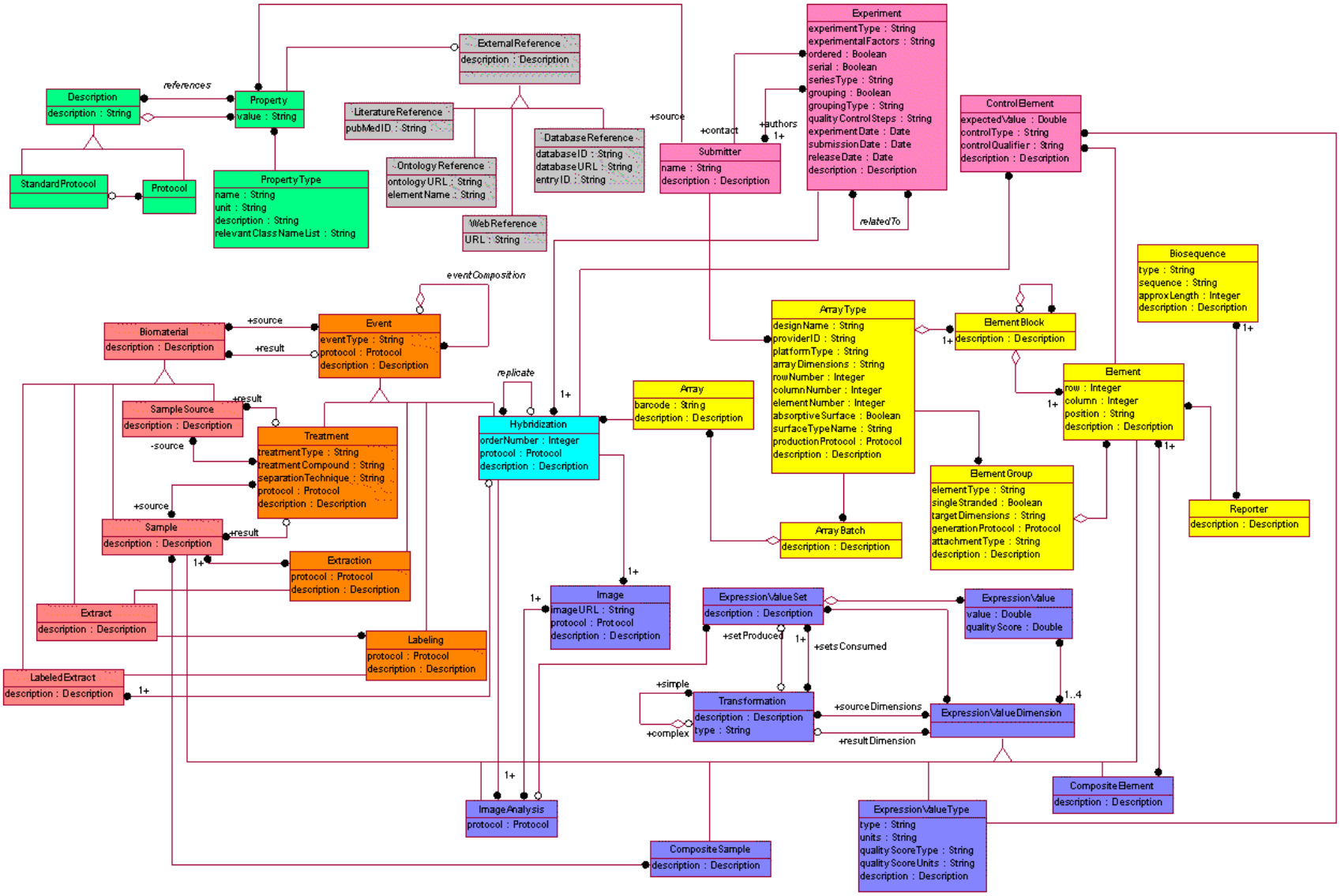


ArrayExpress

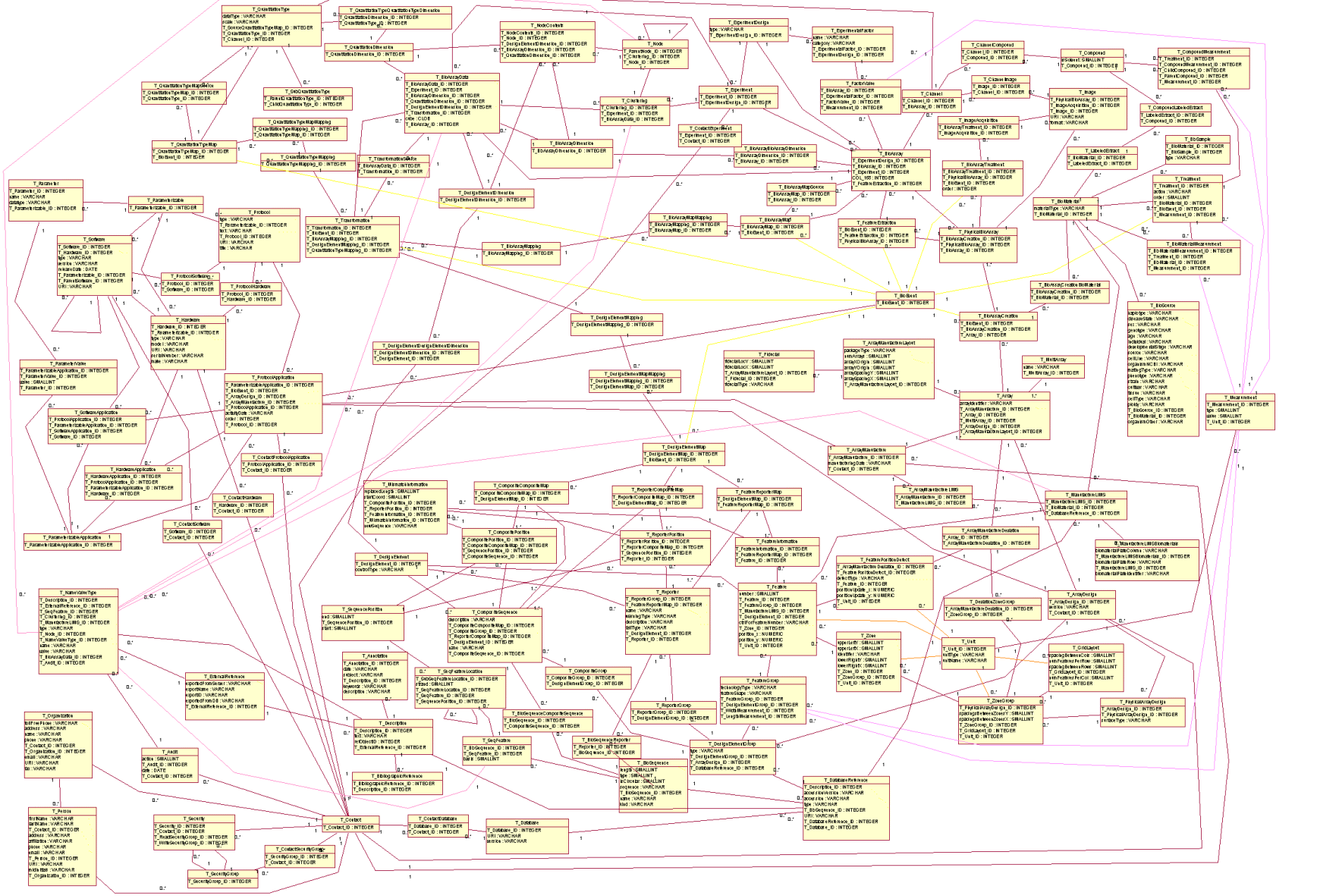
- Is an implementation of the MAGE-OM model
- MAGE-OM has been accepted by the OMG as a biosciences standard
- MAGE-OM is a platform independent model developed in UML



ArrayExpress conceptual model

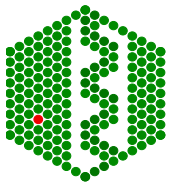


Simplified ArrayExpress model



ArrayExpress details

- Database schema derived from MAGE-OM
- Standard SQL, we use Oracle
- Validating data loader for MAGE-ML - generated
- Web interface (first release 12.2.2002)
  - Queries - experiment, array, sample
  - Browsing – views on expt
- Object model-based query mechanism, automatic mapping to SQL



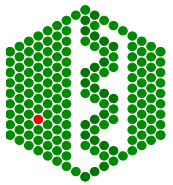
Data in ArrayExpress

Currently-

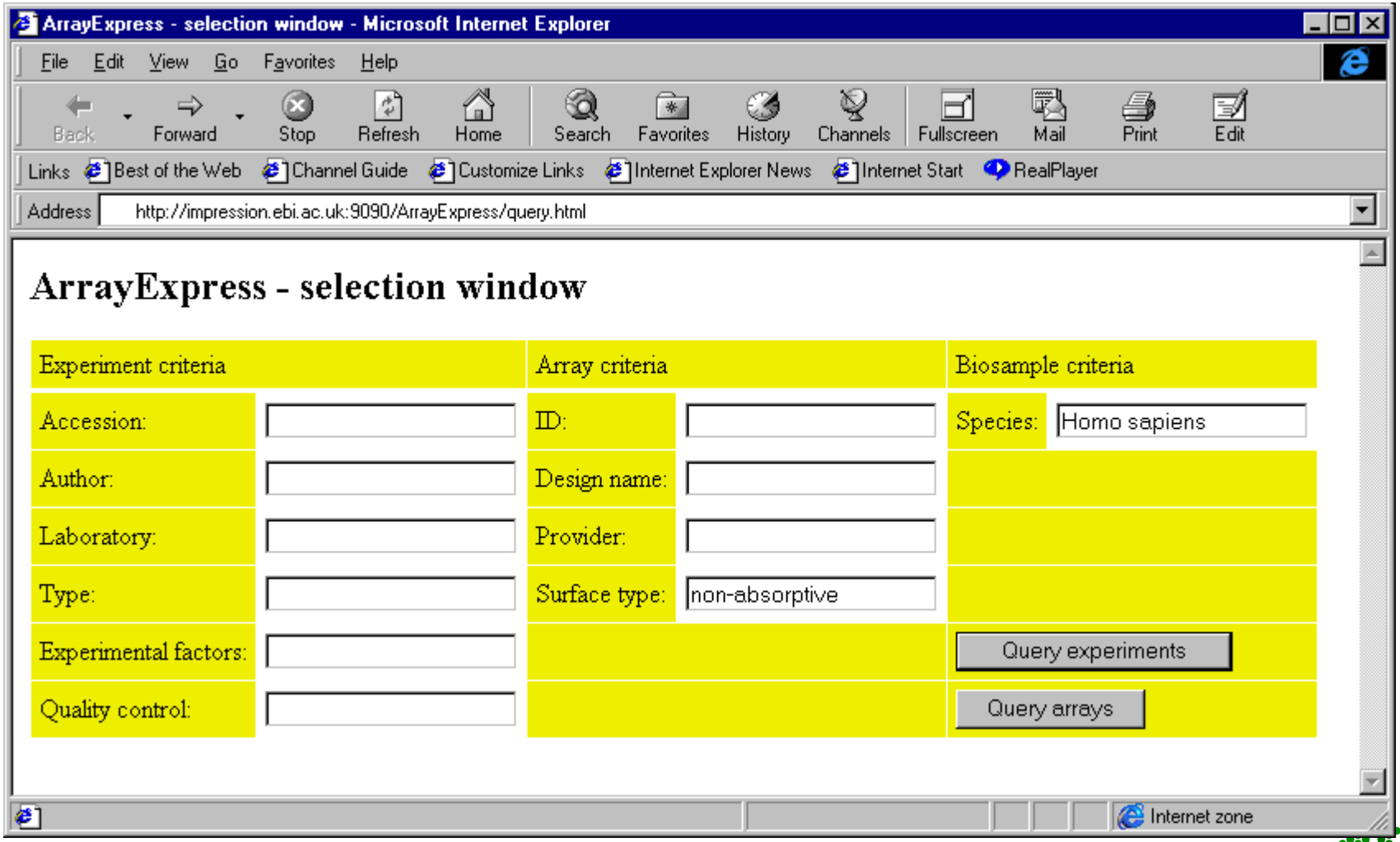
- Human data (ironchip) from EMBL
- Yeast data from EMBL
- S. pombe data Sanger Institute
- Available as example annotated and curated data sets

Near future -

- TIGR array descriptions and data
- Affymetrix array designs
- Direct pipeline from Sanger Institute database
- HGMP mouse data
- EMBL Anopheles data

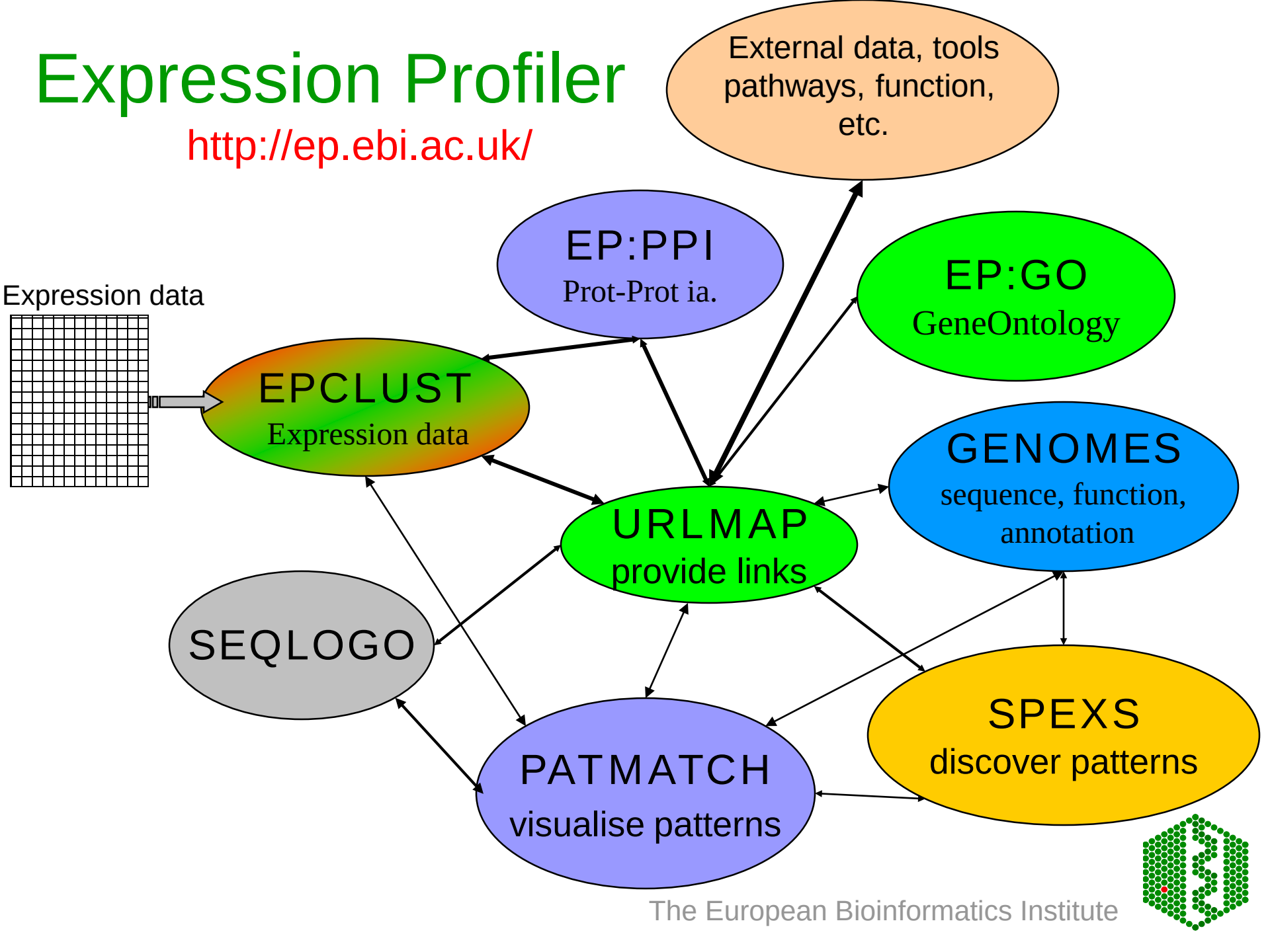


ArrayExpress - queries



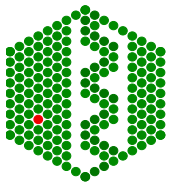
Expression Profiler

http://ep.ebi.ac.uk/



ArrayExpress curation effort

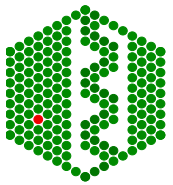
- User support and help documentation
- Curation at source (not destination)
- Support on ontologies and CV's
- Minimize free text, removal of synonyms
- MIAME encouragement
- Help on MAGE-ML
- **Goal:** to provide high-quality, well-annotated data to allow automated data analysis



Data Submission routes

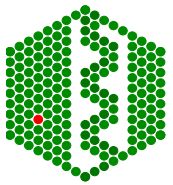
- Via MAGE-ML generated from a local database, (array, protocol and experiment submissions)

- Via MIAMExpress, a MIAME compliant data annotation tool (array, protocol and experiment submissions)



MIAMExpress submission and annotation tool

- Based on MIAME concepts and questionnaire
- Experiment, Array, Protocol submissions
- Uses CV/Ontology wherever possible
- Future versions organism specific pages and related linked ontologies
- Allows user driven ontology development
- Will be developed according to user needs
- Can be used as an update tool
- Can be used as basis of LIMS

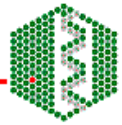


MIAMExpress

EBI Home

Help

EMBL
European Bioinformatics Institute



Experimental Design:

Sample of experiment:

Labeled extract:

How many hybridisation do you want to create (or add) for this experiment? :

1 If 10+ please specify:

Create

List of existing hybridisation(s) related to this experiment:

Hybrid.ID Array design name Array Batch Serian No. Incorporating labeled extracts

HYBRID1SUB3 EMBL Yeast array

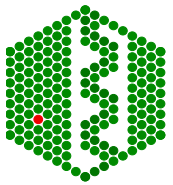
LABEL2EXTRACT2SAMPLE8SUB3
LABEL2EXTRACT2SAMPLE8SUB3
LABEL2EXTRACT2SAMPLE8SUB3
LABEL2EXTRACT2SAMPLE8SUB3
LABEL2EXTRACT2SAMPLE8SUB3

HYBRID2SUB3 EMBL Ironchip

LABEL2EXTRACT2SAMPLE8SUB3
LABEL2EXTRACT2SAMPLE8SUB3
LABEL2EXTRACT2SAMPLE8SUB3
LABEL2EXTRACT2SAMPLE8SUB3
LABEL2EXTRACT2SAMPLE8SUB3

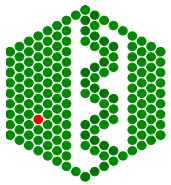
Expected Users

- Users with limited local bioinformatics support
- Users of bought in arrays without LIMS
- Small scale users with self made arrays who will need to provide a description
- Commercial array descriptions will be provided



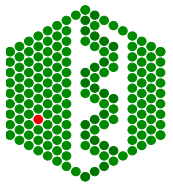
MIAMExpress future developments

- Species and domain specific pages and ontologies, ontology development
- Life-span of data submissions is long
- Integrated curation control, submissions tracking
- Full compatibility with ArrayExpress
- Full MAGE-OM, data updating
- Usability, flexibility, scalability, platform independence
- User needs, free in-house installation



ArrayExpress Future

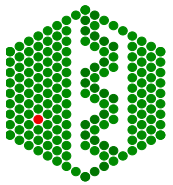
- Loading of public data in MAGE-ML format (TIGR, EMBL, DESPRAD partners) into ArrayExpress
- V2.0 MIAMExpress, the KeyLargoExpress
- Improved query interfaces
- Further ontology development and integration into tools
- Curation tools
- Join MGED www.mged.org



Resources

- Schemas for both ArrayExpress and MIAMExpress, access to code
- Annotation examples in MAGE-ML
- MIAME glossary, MAGE-MIAME-ontology mappings
- List of ontology resources from MGED pages
- MAGE-OM tutorials at MGED meetings
- MAGE-OM support for submitters from EBI
- MAGE-stk API's

www.mged.org www.ebi.ac.uk/microarray



Acknowledgments

- Microarray Informatics Team, EBI
- Chris Stoeckert, U. Penn.
- Members of MGED
- Sanger Institute - Rob Andrews, Jurg Bahler, Adam Butler, Kate Rice,
- EMBL Heidelberg - Wilhelm Ansorge, Martina Muckenthaler,Thomas Preiss