Gene Expression RFP response

Initial Submission

EMBL-EBI (European Bioinformatics Institute)

OMG document # lifesci/2000-11-16

OMG Document lifesci/00-03-09 (Gene Expression RFP)

Version 1.0

20 November 2000

The material in this document is submitted to the OMG for evaluation. Submission of this document does not represent a commitment to implement any portion of this specification in the products of the submitters.

WHILE THE INFORMATION IN THIS PUBLICATION IS BELIEVED TO BE ACCURATE, THE COMPANIES LISTED ABOVE MAKE NO WARRANTY OF ANY KIND WITH REGARD TO THIS MATERIAL INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. The companies listed above shall not be liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance or use of this material. The information contained in this document is subject to change without notice.

1. Preface

This submission is in response to LSR RFP, Gene Expression, Object Management Group (OMG) Document lifesci/00-03-09 (Gene Expression RFP)

1.1 Submission Contact Points

Ugis Sarkans
European Bioinformatics Institute
EMBL Outstation – Hinxton
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD
United Kingdom
(+44) 1223 494603
ugis@ebi.ac.uk

1.2 Supporting Organisations

The proposal is supported by the Microarray Gene Expression Database (MGED) group and has been prepared by the Microarray Markup Language (MAML) working group of MGED.

The MGED group is an open discussion group established at the Microarray Gene Expression Database meeting MGED I on November 16-17, 1999, in Cambridge, UK. 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 normalization methods. The underlying goal is to facilitate the establishing of gene expression data repositories, comparability of gene expression data from different sources and interoperability of different gene expression databases and data analysis software. Since 1999 the group has had two general meetings and the third one is scheduled for March 28-30, 2001, in Stanford US. MGED group includes representatives from the EMBL-EBI, National Center for Biotechnology Information (NCBI), National Center for Genome Research (NCGR), DNA Databank of Japan (DDBJ), National Human Genome Research Institute, German Cancer Research Centre, Stanford University, University of California at Berkeley, University of Colorado, Rockefeller University, Whitehead Institute, Affymetrix, Incyte and Gene Logic Ltd. MGED has established five working groups, including MAML working group, which is coordinated by Paul Spellman from the University of California at Berkeley (UCLB).

For more information on MGED see http://www.mged.org/.

1.3 Acknowledgements

Below is the list of authors from the MGED MAML working group, who have substantially contributed to the proposal:

Paul Spellman	UCLB	spellman@bdgp.lbl.gov
Alvis Brazma	EMBL-EBI	brazma@ebi.ac.uk
Jack Chen	NIH	xchen@helix.nih.gov
Mike Cherry	Stanford University	cherry@stanford.edu
Jonathan Epstein	NIH	jonathan_epstein@nih.gov
Carol Harger	NCGR	cah@ncgr.org
Pascal Hingamp	University Marselle	hingamp@ciml.univ-mrs.fr

Alex Lash	NCBI	alash@ncbi.nlm.hih.gov
Isaac Neuhaus	BMS	isaac.neuhaus@bms.com
John Quackenbush	TIGR	johnq@tigr.org
V. Ravichandran	NIST	vravi@nist.gov
Alan Robinson	EMBL-EBI	alan@ebi.ac.uk
Ugis Sarkans	EMBL-EBI	ugis@ebi.ac.uk
Jason Stuart	Open Informatics	jason_e_stewardt@yahoo.com
Ron Tailor	CU School of Medicine	taylor@uchsc.edu
R. Yang	GCG	yang@gcg.com
Jiaye Zhou	NCGR	JZ@ncgr.org

The authors would like to thank all the MGED members who have contributed to the proposal.

1.4 Proof of Concept

MGED group, which includes representatives from most of the major microarray data providers in academia and industry, and major public bioinformatics databases centres, is committed to establishing standards for gene expression profiling.

The EMBL-EBI, NCBI and NCGR are establishing a public repositories for gene expression data which will use the data format proposed in this document. Although currently the data format is based on XML specification, the complete object description will be added in the next submission.

1.5 Response to RFP Requirements

All the mandatory requirements listed in the items 6.5 of the RFP are fulfilled in this proposal

2. Introduction

We propose a framework for describing information about a DNA-array experiment and a data format – Microarray Markup Language (MAML) – for communicating this information. The information includes details about:

1. Experimental design: the set of the hybridization experiments as a whole;
2. Array design: each array used and each element (spot) on the array;
3. Samples: samples used, the extract preparation and labeling;
4. Hybridizations: procedures and parameters;
5. Measurements: images, quantitation, specifications;
6. Controls: types, values, specifications.

MAML is based on the Extendible Markup Language XML. MAML is independent of the particular experimental platform and provides a framework for describing experiments done on all types of DNA-arrays, including spotted and synthesized arrays, as well as oligo-nucleotide and cDNA arrays, and is independent of the particular image analysis and data normalization methods. MAML does not impose any particular image analysis or data normalization method, but instead provides format to represent microarray data in a flexible way, which allows to represent data obtained from not only any existing microarray platforms, but also many of the possible future variants, including protein arrays. The format allows representation of raw and processed microarray data. The format is compatible with the definition of the “minimum information about a microarray experiment” (MIAME) proposed by the MGED group, see http://www.mged.org/.

The MGED group is an open discussion group initially established at the Microarray Gene Expression Database meeting MGED 1 (November, 1999, Cambridge, UK). 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 normalization methods. The underlying goal is to facilitate the establishing of gene expression data repositories, comparability of gene expression data from different sources and interoperability of different gene expression databases and data analysis software.

In the next two sections, we describe the MIAME standard, which describes the content of the information that has to be represented by a data format for microarray gene expression data representation (according to MGED recommendations), followed by the MAML DTD, which defines the actual XML based data format.

3. Minimum information about a microarray experiment - (MIAMI)

Endorsed by MGED steering committee meeting November 17, 2000

The goal of the MIAME is to specify the minimum information that must be reported about a microarray based gene expression monitoring experiment in order to ensure the interpretability of the results and their reproducibility by third parties. The background aim is to help establishing public repositories and data exchange format for microarray based gene expression data. Scientific journals will be encouraged to adopt editorial policies requiring data submissions to repositories, once MIAMI compliant repositories are established.

Introduction:

The definition of the minimum information is aimed at cooperative data providers, and not as a legal document meant to close possible loopholes in not providing the information.

Among the concepts in the definition is a list of "qualifier, value, source" triplets, where the "source" is either user defined, or a reference to an externally defined ontology or controlled vocabulary, such as the species taxonomy database at NCBI. Where necessary, the authors are encouraged to define their own qualifiers and provide the appropriate values so that the list as the whole gives sufficient information to interpret the particular part of the experiment. The judgement regarding the necessary level of detail is left to the submitters themselves. In future these `voluntary' qualifier lists may be gradually substituted by required fields, as the respective ontologies are developed.

Parts of the MIAME can be provided as a reference or link to an externally existing description. For instance, for commercial or other standard arrays all the required information should be normally provided only once by the array provider and referenced by the users. Standard protocols should also normally be provided only once.

Definition:

The minimum information about a published microarray based gene expression experiment should include the description of

1. Experimental design: the set of the hybridisation experiments as a whole
2. Array design: each array used and each element (spot) on the array
3. Samples: samples used, the extract preparation and labeling

4. Hybridisations: procedures and parameters
5. Measurements: images, quantitation, specifications
6. Controls: types, values, specifications

The following details should be provided for each array, each sample, hybridisation and measurement in the experiment set:

1. Experimental design: the set of the hybridisation experiments as a whole

a) author (submitter), laboratory, contact information, links (URL)

b) type of the experiment - maximum one line for instance:

- normal vs. diseased comparison
- treated vs. untreated comparison
- time course
- dose response
- effect of gene knock-out
- effect of gene knock-in (transgenics)
- shock

(multiple types possible)

c) experimental factors (e.g., time, dose, genetic variation),

d) the list of platforms used,

e) single or multiple hybridisations,

For multiple hybridisations:

- ordered/unordered
- serial (yes/no)
- type (e.g., time course, dose response)
- grouping (yes/no)
- type (e.g., normal vs. diseased, multiple tissue comparison)
- list of the samples and arrays used in the experiment and description of the relationship between them: each sample and each array should be assigned a unique id in the experiment set and all the relationships should be listed with appropriate comments
- which hybridisations are replicates

f) quality related indicators

- does a related peer-reviewed publication exist
- number of replicate hybridisations
- any other quality control steps taken (polya, unspecific binding etc.)

g) optional user defined "qualifier, value, source" list (see Introduction)

h) a free text description of the experiment set or a link to a publication

2. Array design: each array used and each element (spot) on the array.

a) array

- array design name (e.g., "Stanford Human 10K set")
- platform type: insitu synthesized or spotted

- provider (source)
- surface type: absortive/nonabsortive
- surface type name
- array dimensions
- number of elements on the array
- a reference system allowing to locate each element (spot) on the array (in the simplest case the number of columns and rows is sufficient)
- unique ID from the provider
- production protocol (obligatory if applicable)
- optional "qualifier, value, source" list (see Introduction)

b) element (spot) on the array - elements may be simple, i.e., containing only identical molecules, or composite, i.e., containing different oligonucleotides obtained from the same reference molecule; for each element the following must be given:

- position on the array allowing to identify the spot in the image (see 5. a) below);
- element type: synthesized oligo-nucleotides, PCR products, plasmids, colonies, other;
- clone information, obligatory for elements obtained from clones:
  - clone ID, clone provider, date, availability
- sequence information, obligatory for synthetic elements:
  - sequence accession number in DDBJ/EMBL/GenBank if known
  - sequence itself (if databases do not contain it)
  - number of oligos and the reference sequence (or accession number) for multiple oligo-per-element type chips, plus the
  - oligo-sequences, if given
- approximate lengths if exact sequence not known
- singe or double stranded
- element (spot) dimensions
- element generation protocol that includes sufficient information to reproduce the element;
- gene name and links to appropriate databases (e.g., SWISS-PROT, or organism specific databases), if known and relevant
- if the element can be used for normalization or control (e.g., element should have expected value)

3. Samples: samples used, extract preparation and labeling

a) sample source and treatment:

- organism (NCBI taxonomy)
- additional "qualifier, value, source" list; each qualifier in the list is obligatory if applicable; the list includes:
  - cell source and type (if derived from primary sources (s))
  - sex
  - age
  - development stage
  - organism part (tissue)
  - animal/plant strain or line
  - genetic variation (e.g., gene knockout, transgenic variation)
  - individual

    - individual genetic characteristics (e.g., disease alleles, polymorphisms)
    - disease state or normal
    - target cell type
    - cell line and source (if applicable)
    - in vivo treatments (organism or individual treatments)
    - in vitro treatments (cell culture conditions)
    - treatment type (e.g., small molecule, heat shock, cold shock, food deprivation)
    - compound
    - separation technique (e.g., none, trimming, microdissection, FACS)
- laboratory protocol for sample treatment

b) hybridisation extract preparation

- laboratory protocol for extract preparation, including:
    - extraction method
    - whether total RNA, mRNA, or genomic DNA is extracted
    - amplification (RNA polymerases, PCR)
- optional "qualifier, value, source" list (see Introduction)

c) labeling

- laboratory protocol for labelling, including:
    - amount of nucleic acids labeled
    - exogenous sequences (spikes) added
    - label used (e.g., Cy3, Cy5, 33P)
- optional "qualifier, value, source" list (see Introduction)

4. Hybridisations: procedures and parameters

- laboratory protocol for hybridisation, including:
    - the solution (e.g., concentration of solutes)
    - blocking agent
    - wash procedure
    - quantity of labelled target used
    - time, concentration, volume, temperature
    - description of the hybridisation instruments
- optional "qualifier, value, source" list (see Introduction)

5. Measurements: images, quantitation, specifications:

a) hybridisation scan raw data:

a1) the scanner image file (e.g., TIFF) from the hybridised microarray scanning;

a2) scanning information:

- parsed header of the TIFF file, including laser power, spatial resolution, pixel space, PMT voltage;
- laboratory protocol for scanning, including:
    - scanning hardware

  - scanning software

b) image analysis and quantitation

b1) the complete image analysis output (of the particular image analysis software) for each element (or composit element - see 2.b)), for each channel;

b2) image analysis information:

- image analysis software specification and version, availability, and the description of the algorithm
- all parameters

c) summarized information from possible replicates

c1) derived measurement value summarizing related elements as used by the author (this may constitute replicates of the element on the same or different arrays or hybridisations, as well as different elements related to the same entity e.g., gene)

c2) reliability indicator for the value of c1) as used by the author (e.g., standard deviation); may be "unknown"

c3) specification how c1 and c2 are calculated; the specification should be bases on b1

6. Normalisation controls, values, specifications for hybridisations

a) Normalization strategy
- spiking
- “housekeeping gene”
- total array
- optional used defined “quality value”

b) Normalisation algorithm
- linear regression
- log-linear regression
- ratio statistics
- log(ratio) mean/median centering
- nonlinear regression
- optional used defined “quality value”

c) Control array elements
- position (the abstract coordinate on the array)
- control type (spiking, normalization, negative, positive)
- control qualifier (endogenous, exogenous)
- optional used defined “quality value”

d) Hybridisation extract preparation
- spike type
- spike qualifier
- target element
- optional used defined “quality value”

4. MAML DTD

```
<!-------------------------------------------------------------------->
<!-- MAML DOCUMENT CLUSTER                                          -->

<!ELEMENT  maml          (analysis_list?,
                         array_platform_list?,
                         contact_list?,
                         creation_info,
                         data_set_list?,
                         experiment_set_list?,
                         hardware_list?,
                         protocol_list?,
                         sample_list?,
                         software_list?,
                         publication_list) >


<!-------------------------------------------------------------------->
<!-- CREATION INFORMATION                                           -->
<!--                    A description of the creator of the XML
                        document (human, software, hardware)        -->

<!ELEMENT  creation_info EMPTY >

<!--                    date is an ISO date string                  -->
<!ATTLIST  creation_info
              date        CDATA                             #REQUIRED
              contact_id  IDREF                             #IMPLIED
              software_id IDREF                             #IMPLIED  >


<!--                    Contact can specify either an individual
                        researcher or an organization               -->
<!ELEMENT  contact_list  (contact+) >
<!ELEMENT  protocol_list (protocol+) >
<!ELEMENT  hardware_list (hardware+) >
<!ELEMENT  software_list (software+) >

<!ELEMENT  contact       (parameter*) >
<!ATTLIST  contact
              id             ID                             #REQUIRED
              last_name      CDATA                          #IMPLIED
              first_name     CDATA                          #IMPLIED
              middle_name    CDATA                          #IMPLIED
              type           CDATA                          #IMPLIED
              lab            CDATA                          #IMPLIED
              department     CDATA                          #IMPLIED
              organization   CDATA                          #IMPLIED
              street         CDATA                          #IMPLIED
              city           CDATA                          #IMPLIED
              province_state CDATA                          #IMPLIED
              country        CDATA                          #IMPLIED
              postal_code    CDATA                          #IMPLIED
              phone          CDATA                          #IMPLIED
              fax            CDATA                          #IMPLIED
              email          CDATA                          #IMPLIED
              uri            CDATA                          #IMPLIED  >
```

```
<!--                    Can represent PCR, scanner, array printer,
                        etc.                                        -->
<!ELEMENT  hardware      (description?,
                         parameter*) >
<!ATTLIST  hardware
              id            ID                              #REQUIRED
              contact_id    IDREF                           #IMPLIED
              type          CDATA                           #IMPLIED
              make          CDATA                           #IMPLIED
              model         CDATA                           #IMPLIED
              serial_number CDATA                           #IMPLIED
              year          CDATA                           #IMPLIED
              uri           CDATA                           #IMPLIED  >


<!ELEMENT  software      (description?,
                         parameter*) >
<!ATTLIST  software
              id               ID                           #REQUIRED
              contact_id       IDREF                        #IMPLIED
              hardware_ids     IDREFS                       #IMPLIED
              type             CDATA                        #IMPLIED
              name             CDATA                        #REQUIRED
              version          CDATA                        #IMPLIED
              year             CDATA                        #IMPLIED
              operating_system CDATA                        #IMPLIED
              uri              CDATA                        #IMPLIED  >


<!-- III     Protocols                                              -->
<!ELEMENT  protocol     (standard_protocol,
                         db_xref?,
                         protocol_deviations?,
                         protocol_abstract?) >
<!ATTLIST  protocol
              id        ID                                  #REQUIRED
              name      CDATA                               #IMPLIED
              type      CDATA                               #IMPLIED  >

<!ELEMENT  protocol_abstract (#PCDATA) >
<!ATTLIST  protocol_abstract
              xml:space preserve                            #FIXED    >

<!ELEMENT  standard_protocol (#PCDATA) >
<!ATTLIST  standard_protocol
              xml:space preserve                            #FIXED    >

<!ELEMENT  protocol_deviations (#PCDATA) >
<!ATTLIST  protocol_deviations
              xml:space preserve                            #FIXED    >


<!-- IV      Data                                                   -->
<!ELEMENT  data_set_list (data_set+) >

<!--                    for each grouping the first item in the
                        pair is the rows of the matrix, and the
                        second element is columns                   -->
<!ELEMENT  data_set      ((matrix_axes,
                          matrix_data),
                         (tagged_data_internal|
                          tagged_data_external)) >
<!ATTLIST  data_set
              id          ID                                 #REQUIRED
              name        CDATA                              #IMPLIED
              description CDATA                              #IMPLIED  >

<!ELEMENT  matrix_data  (ascii_data_internal |
                         ascii_data_external |
                         binary_data_external)*                        >

<!ELEMENT  matrix_axes  (matrix_row_list,
                         matrix_column_list,
                         matrix_stack) >
<!ELEMENT  matrix_row_list (matrix_row+) >
<!ELEMENT  matrix_row   EMPTY >
<!ATTLIST  matrix_row
              element_id      IDREF                          #IMPLIED
              image_id        IDREF                          #IMPLIED
              quantitation_id IDREF                          #IMPLIED  >

<!ELEMENT  matrix_column_list (matrix_column+) >
<!ELEMENT  matrix_column EMPTY >
<!ATTLIST  matrix_column
             element_id      IDREF                           #IMPLIED
             image_id        IDREF                           #IMPLIED
             quantitation_id IDREF                           #IMPLIED  >

<!ELEMENT  matrix_stack (matrix+) >
<!ELEMENT  matrix       EMPTY >
<!ATTLIST  matrix
             element_id      IDREF                           #IMPLIED
             image_id        IDREF                           #IMPLIED
             quantitation_id IDREF                           #IMPLIED  >

<!--                    axis_key refers to either an <element>:id or
                        <composite_element>:id; or an <image>:id or
                        <composite_image>:id; or a <quantitation>:id or
                        <composite_quantitation>:id this is intended to
                        reference the missing third dimension of the data
                        matrix                                      -->
<!--
Data stored internally should be treated as a white
space delimited matrix where null values are specified
as 'NULL'. Carriage returns to delineate the ends of
rows are not necessary.
 -->
<!ELEMENT  ascii_data_internal (#PCDATA) >
<!ATTLIST  ascii_data_internal
              id         ID                                  #REQUIRED
              type       CDATA                               #REQUIRED
              derivation CDATA                               #REQUIRED >


<!--
Data stored externally should be treated as a white
space delimited matrix where null values are specified
as 'NULL'. Carriage returns to delineate the ends of
rows are not necessary.
 -->
<!ELEMENT  ascii_data_external EMPTY >
```

```
<!ATTLIST  ascii_data_external
              id        ID                                  #REQUIRED
              type      CDATA                               #REQUIRED
              file_uri  CDATA                               #REQUIRED >

<!ELEMENT  tagged_data_internal (tagged_data+) >
<!ATTLIST  tagged_data_internal
              id        ID                                  #REQUIRED
              type      CDATA                               #REQUIRED >


<!ELEMENT  tagged_data_external (tagged_data+) >
<!ATTLIST  tagged_data_external
              id        ID                                  #REQUIRED
              type      CDATA                               #REQUIRED >


<!ELEMENT  tagged_data EMPTY >
<!ATTLIST  tagged_data
              element_id      IDREF                         #REQUIRED
              image_id        IDREF                         #REQUIRED
              quantitation_id IDREF                         #REQUIRED
              data            CDATA                         #REQUIRED >

<!ELEMENT  binary_data_external EMPTY >
<!ATTLIST  binary_data_external
              id        ID                                  #REQUIRED
              axis_key  IDREF                               #REQUIRED
              type      CDATA                               #REQUIRED
              file_uri  CDATA                               #REQUIRED >

<!ELEMENT  parameter EMPTY >
<!ATTLIST  parameter
              name      CDATA                               #REQUIRED
              value     CDATA                               #REQUIRED >

<!-- V       Analysis -->
<!ELEMENT  analysis_list  (analysis+) >
<!ELEMENT  analysis (quantitation_list,
                    composite_image_list,
                    composite_quantitation_list,
                    composite_element_list) >

<!--                    For primary <quantitation> the 'name'
                        should be the same as the column name
                        provided by the scanner software           -->
<!--                    For a primary <quantitation> the
                        'software_id' really ought to be required
                        and should refer to the scanner software
                        that produced the data                     -->
<!ELEMENT  quantitation_list (quantitation+) >
<!ELEMENT  quantitation EMPTY >
<!ATTLIST  quantitation
              id          ID                                #REQUIRED
              name        CDATA                             #REQUIRED
              software_id IDREF                             #IMPLIED
              protocol_id IDREF                             #IMPLIED  >

<!ELEMENT  composite_element_list (composite_element+) >
<!ELEMENT  composite_element EMPTY >
<!ATTLIST  composite_element
```

```
              id          ID                                #REQUIRED
              element_ids IDREFS                            #REQUIRED >


<!ELEMENT  composite_image_list (composite_image+) >
<!ELEMENT  composite_image EMPTY >
<!--                    protocol_id references a protocol that
                        describes the method used to create the
                        composite elements from the primary
                        measurements                               -->
<!ATTLIST  composite_image
              id          ID                                #REQUIRED
              image_ids   IDREFS                            #REQUIRED
              protocol_id IDREF                             #IMPLIED
              software_id IDREF                             #IMPLIED  >

<!ELEMENT  composite_quantitation_list (composite_quantitation)+>

<!ELEMENT  composite_quantitation EMPTY >
<!--                    protocol_id references a protocol that
                        describes the method used to create the
                        composite  from the primary
                        measurements                               -->
<!--                    We're not sure that software_id is useful
                        in this context                            -->
<!ATTLIST  composite_quantitation
             id              ID                             #REQUIRED
             quantitaion_ids IDREFS                         #REQUIRED
             protocol_id     IDREF                          #IMPLIED
             software_id     IDREF                          #IMPLIED  >

<!-- VIII    Experiment Set -->
<!ELEMENT  experiment_set_list (experiment_set+) >
<!ELEMENT  experiment_set (experimental_design,
                           extract_list,
                           hybridization_list,
                           control_element_list,
                           labeled_extract_list,
                           sample_list) >
<!ATTLIST  experiment_set
              local_accession_number CDATA                  #IMPLIED
              experiment_type        CDATA                  #IMPLIED
              publication_id         IDREF                  #IMPLIED
              contact_id             CDATA                  #IMPLIED
              submission_date        CDATA                  #IMPLIED
              release_date           CDATA                  #IMPLIED
              experiment_date        CDATA                  #IMPLIED  >

<!ELEMENT  experimental_design (biology_description,
                                analysis_description,
                                experimental_factors,
                                quality) >

<!ELEMENT  biology_description (#PCDATA) >
<!ATTLIST  biology_description
              xml:space preserved                           #FIXED >

<!ELEMENT  analysis_description (#PCDATA) >
<!ATTLIST  analysis_description
              xml:space preserved                           #FIXED >
```

```
<!ELEMENT  experimental_factors (#PCDATA) >
<!ATTLIST  experimental_factors
              xml:space preserved                           #FIXED >


<!--                    We understand that this is limited and
                        insufficient, but we believe that quality
                        control is an important issue              -->
<!ELEMENT  quality (replicates, quality_info*) >
<!ATTLIST  quality
              has_replicates (true|false)                    #REQUIRED >
              peer_reviewed  (true|false)                    false     >

<!ELEMENT  replicates (description?) >

<!ELEMENT  quality_info (#PCDATA) >

<!ELEMENT  control_element_list (control_element+) >
<!ELEMENT  control_element EMPTY >
<!ATTLIST  control_element
             id              ID                             #REQUIRED
             expected_value  CDATA                          #REQUIRED
             quantitation_id IDREF                          #REQUIRED
             element_id      IDREF                          #REQUIRED >


<!ELEMENT  hybridization_list (hybridization+) >
<!ELEMENT  hybridization (image+) >
<!ATTLIST  hybridization
             name                CDATA                      #REQUIRED
             protocol_ids        IDREFS                     #IMPLIED
             labeled_extract_ids IDREFS                     #REQUIRED
             control_element_ids IDREFS                     #REQUIRED
             array_id            IDREF                      #REQUIRED
             id                  ID                         #REQUIRED >


<!ELEMENT  image EMPTY >
<!ATTLIST  image
              protocol_id         IDREF                     #REQUIRED
              labeled_extract_ids IDREFS                    #REQUIRED
              software_id         IDREF                     #REQUIRED
              hardware_id         IDREF                     #REQUIRED
              file_uri            CDATA                     #REQUIRED
              file_header         CDATA                     #IMPLIED
              microns_per_pixel   CDATA                     #IMPLIED
              image_identifier    CDATA                     #REQUIRED >


<!ELEMENT  sample_list   (primary_sample|
                          derived_sample)+ >

<!ELEMENT  derived_sample (treatment+) >
<!ATTLIST  derived_sample
              id                ID                          #REQUIRED
              parent_sample_ids IDREFS                      #REQUIRED >

<!ELEMENT  treatment (measurement) >
<!ATTLIST  treatment
              action      CDATA                             #REQUIRED
              object      CDATA                             #IMPLIED
              protocol_id IDREF                             #IMPLIED
              order       CDATA                             #REQUIRED >


<!--   We don't yet have a full ontology so the primary sample
       should include the following kinds of values:

       organism_ncbi
       organism_other
       cell_source
       cell_type
       sex
       age
       development_stage
       organism_part (tissue)
       strain_or_line
       genetic_variation
       individual
       genotype
       disease_state
       target_cell_type
       cell_line_and_source
       in_vivo_treatments
       in_vitro_treatments
       treatment_type
       compound
       separation_technique             -->
<!ELEMENT  primary_sample (parameter|generic_measure)* >
<!ATTLIST  primary_sample
              id        ID                                  #REQUIRED >

<!ELEMENT  extract_list (extract+) >
<!ELEMENT  extract (description?,parameter*) >
<!ATTLIST  extract
              id          ID                                #REQUIRED
              protocol_id IDREF                             #REQUIRED
              type        (total_rna|mrna|dna)              #REQUIRED
              sample_ids  IDREFS                            #REQUIRED
              label_name  CDATA                             #REQUIRED
              name        CDATA                             #IMPLIED  >

<!ELEMENT  labeled_extract_list (labeled_extract+) >

<!ELEMENT  labeled_extract (description?,parameter*) >
<!ATTLIST  labeled_extract
              id          ID                                #REQUIRED
              protocol_id IDREF                             #REQUIRED
              extract_ids IDREFS                            #REQUIRED
              name        CDATA                             #IMPLIED  >

<!-------------------------------------------------------------------->
<!-- DESCRIPTIONS                                                   -->

<!ELEMENT  description  CDATA                                         >


<!-------------------------------------------------------------------->
<!-- MEASUREMENT CLUSTER                                            -->

<!ELEMENT  time         EMPTY >
```

```
<!ATTLIST  time
              value      CDATA                              #REQUIRED
              unit       (years |
                          months |
                          weeks |
                          d |
                          h |
                          m |
                          s |
                          ms |
                          us |
                          other)                            #REQUIRED
              other_unit CDATA                              #IMPLIED   >


<!ELEMENT  vector       (distance+) >
<!ELEMENT  distance     EMPTY >
<!ATTLIST  distance
              value      CDATA                              #REQUIRED
              unit       (fm |
                          pm |
                          nm |
                          um |
                          mm |
                          cm |
                          m |
                          other)                            #REQUIRED
              other_unit CDATA                              #IMPLIED   >


<!ELEMENT  temperature  EMPTY >
<!ATTLIST  temperature
              value     CDATA                               #REQUIRED
              unit      (C|F)                               #REQUIRED >

<!ELEMENT  mass         EMPTY >
<!ATTLIST  mass
              value      CDATA                              #REQUIRED
              unit       (kg |
                          g |
                          mg |
                          ug |
                          ng |
                          pg |
                          fg |
                          other)                            #REQUIRED
              other_unit CDATA                              #IMPLIED  >

<!ELEMENT  volume        EMPTY >
<!ATTLIST  volume
              value     CDATA                               #REQUIRED
              unit      (mL |
                         cc |
                         dL |
                         L  |
                         uL |
                         nL |
                         pL |
                         fL |
                         other)                             #REQUIRED
              other_unit CDATA                              #IMPLIED   >
```

```
<!ELEMENT  concentration EMPTY >
<!ATTLIST  concentration
              value      CDATA                              #REQUIRED
              unit       (M |
                          mM |
                          uM |
                          nM |
                          pM |
                          fM |
                          mg_per_mL |
                          mL_per_L |
                          g_per_L |
                          g_percent |
                          other)                            #REQUIRED
              other_unit CDATA                              #IMPLIED   >

<!ELEMENT  quantity     EMPTY >
<!ATTLIST  quantity
              value      CDATA                              #REQUIRED
              unit       ( mol|
                          amol|
                          fmol|
                          pmol|
                          nmol|
                          umol|
                          mmol|
                          molecule)                         #REQUIRED >

<!ELEMENT  generic_measure EMPTY >
<!ATTLIST  generic_measure
              name      CDATA                               #REQUIRED
              value     CDATA                               #REQUIRED
              unit      CDATA                               #REQUIRED >


<!ELEMENT  measurement  (time |
                         distance |
                         vector |
                         quantity |
                         temperature |
                         mass |
                         volume |
                         concentration |
                         generic_measure) >
<!ATTLIST  measurement
              type      (absolute | change)                 #IMPLIED  >



<!--------------------------------------------------------------------->
<!-- RELATIONSHIPS                                                  -->

<!ELEMENT  reference    (db_xref*,description?) >


<!--                    Date is an ISO date string, and is
                        intended to be used to specify the date
                        that the reference was made, not the date
                        the database was released                   -->
<!ELEMENT  db_xref      (parameter*) >
```

```
<!ATTLIST  db_xref
              database          CDATA                       #IMPLIED
              database_version  CDATA                       #IMPLIED
              date              CDATA                       #IMPLIED
              accession         CDATA                       #IMPLIED
              accession_version CDATA                       #IMPLIED
              uri               CDATA                       #IMPLIED  >

<!-------------------------------------------------------------------->
<!-- PUBLICATION                                                   -->

<!ELEMENT  publication_list (publication+) >
<!ELEMENT  publication  (citation | reference) >
<!ATTLIST  publication
              id        ID                                  #REQUIRED >

<!ELEMENT  citation     (abstract?) >
<!ATTLIST  citation
              journal   CDATA                               #IMPLIED
              year      CDATA                               #IMPLIED
              volume    CDATA                               #IMPLIED
              issue     CDATA                               #IMPLIED
              page      CDATA                               #IMPLIED
              authors   CDATA                               #IMPLIED
              publisher CDATA                               #IMPLIED
              editor    CDATA                               #IMPLIED
              uri       CDATA                               #IMPLIED  >

<!ELEMENT  abstract     (#PCDATA) >

<!-------------------------------------------------------------------->
<!-- ARRAY PLATFORM                                                -->
<!--   changes:                                                    -->
<!--      1) 'array_def': exchanged 'type' attribute with          -->
<!--         'surface_type' and 'reporter_type'                    -->
<!--      2) 'reporter' element converted into Paul's              -->
<!--         suggested 'element' element                           -->

<!ELEMENT  array        (description?) >
<!ATTLIST  array
              id                ID                          #REQUIRED
              name              CDATA                       #REQUIRED
              array_platform_id IDREF                       #REQUIRED >

<!ELEMENT  array_platform_list (array_platform|
                                array)+ >

<!ELEMENT  array_platform (array_def) >

<!ATTLIST  array_platform
              id        ID                                  #REQUIRED >


<!ELEMENT  array_def    (description?,
                         reference*,
                         parameter*,
                         element*) >

<!ATTLIST  array_def
              name               CDATA                      #REQUIRED
              contact_id         CDATA                      #REQUIRED
              protocol_id        CDATA                      #REQUIRED
              in_situ_synthesis  (true|false)               #REQUIRED
              spotted            (true|false)               #REQUIRED
              surface_type       (non-absorptive|
                                  absorptive)               #REQUIRED
              surface_type_name  CDATA                      #REQUIRED
              other_surface_type CDATA                      #IMPLIED
              number_of_elements CDATA                      #IMPLIED
              short_axis_length  CDATA                      #IMPLIED
              long_axis_length   CDATA                      #IMPLIED
              element_type       (single-multimer |
                                  multiple-oligomer |
                                  other)                    #REQUIRED
              model_name         CDATA                      #IMPLIED
              version            CDATA                      #IMPLIED
              uri                CDATA                      #IMPLIED  >

<!-------------------------------------------------------------------->
<!-- ELEMENT                                                        -->
<!ELEMENT  element     ((bio_seq|
                         ref_bio_seq|
                         ref_clone)+,
                        gene*,
                        parameter*,
                        measurement*,
                        description?) >

<!--                    sequence_length can be approximate         -->
<!--                    diameter can be approximate                -->
<!--                    empty elements should have an empty
                        <bio_seq>                                  -->
<!ATTLIST  element
              id                ID                          #REQUIRED
              attachment_method CDATA                       #IMPLIED
              strandedness      (single|double)             #IMPLIED
              type              (empty|
                                 pcr|
                                 synthesized_oligo|
                                 intact_plasmid|
                                 colony)                    #REQUIRED
              diameter          CDATA                       #IMPLIED
              sequence_length   CDATA                       #IMPLIED
              location          CDATA                       #IMPLIED
              protocol_id       CDATA                       #IMPLIED
              row               CDATA                       #IMPLIED
              column            CDATA                       #IMPLIED
              block             CDATA                       #IMPLIED
              x_microns         CDATA                       #IMPLIED
              y_microns         CDATA                       #IMPLIED
              name              CDATA                       #IMPLIED  >

<!ELEMENT  bio_seq      (#PCDATA|db_xref) >
<!ELEMENT  ref_bio_seq  (#PCDATA|db_xref) >
<!ELEMENT  gene         (#PCDATA|db_xref) >
<!ELEMENT  ref_clone    (#PCDATA|db_xref) >
```