XML schema for the trait, genotype and mRNA expression data.

Ilze Druka (1), Andrew Wakelin (1), Victor Bassilious (1), Arnis Druka (2).

(1) School of Computing and Creative Technologies, University of Abertay, Dundee DD1 1HG, SCOTLAND United Kingdom.

(2) Genetics Programme, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, SCOTLAND United Kingdom.

Introduction.

The widespread use of high-throughput technologies in molecular biology that generates large data sets has created a need for flexible, easy to use data exchange. This would greatly facilitate early communication of results between researchers and provide vital supplement that can be corrected and adjusted, to the publications. Traditionally spreadsheets like Excel or QuatroPro were used for this purpose but working with large data sets (12 MB) the performance of these applications deteriorates or they cannot accommodate the large data sets at all.

This problem has been solved partially by the serious and successful efforts to manage and make available large quantities of biology related data on the web. Sequence information, protein structures, gene expression are just a few success stories of public depositories alleviating the data sharing and communication as well as providing acknowledged research tools. These are large applications, maintained by teams of developers. The use of the depositories, however is not without difficulties. The data may be either produced by a different technical procedure or processed in a novel way that is different from what the depositories can accommodate. Therefore there is a need for a technology that would allow converting data formats and consolidating data in various ways that are flexible and can be automated.

While traditionally, Perl applications are used for this purpose, the widespread popularity and accolades of XML suggest that it might be one of the technologies to solve this problem. Also the successful use of XML for similar tasks in business, medicine and other fields of biology suggests that it is indeed a suitable technology for flexible and convenient exchange of large data sets.

e-QTL - expression Quantitative Trait Locus

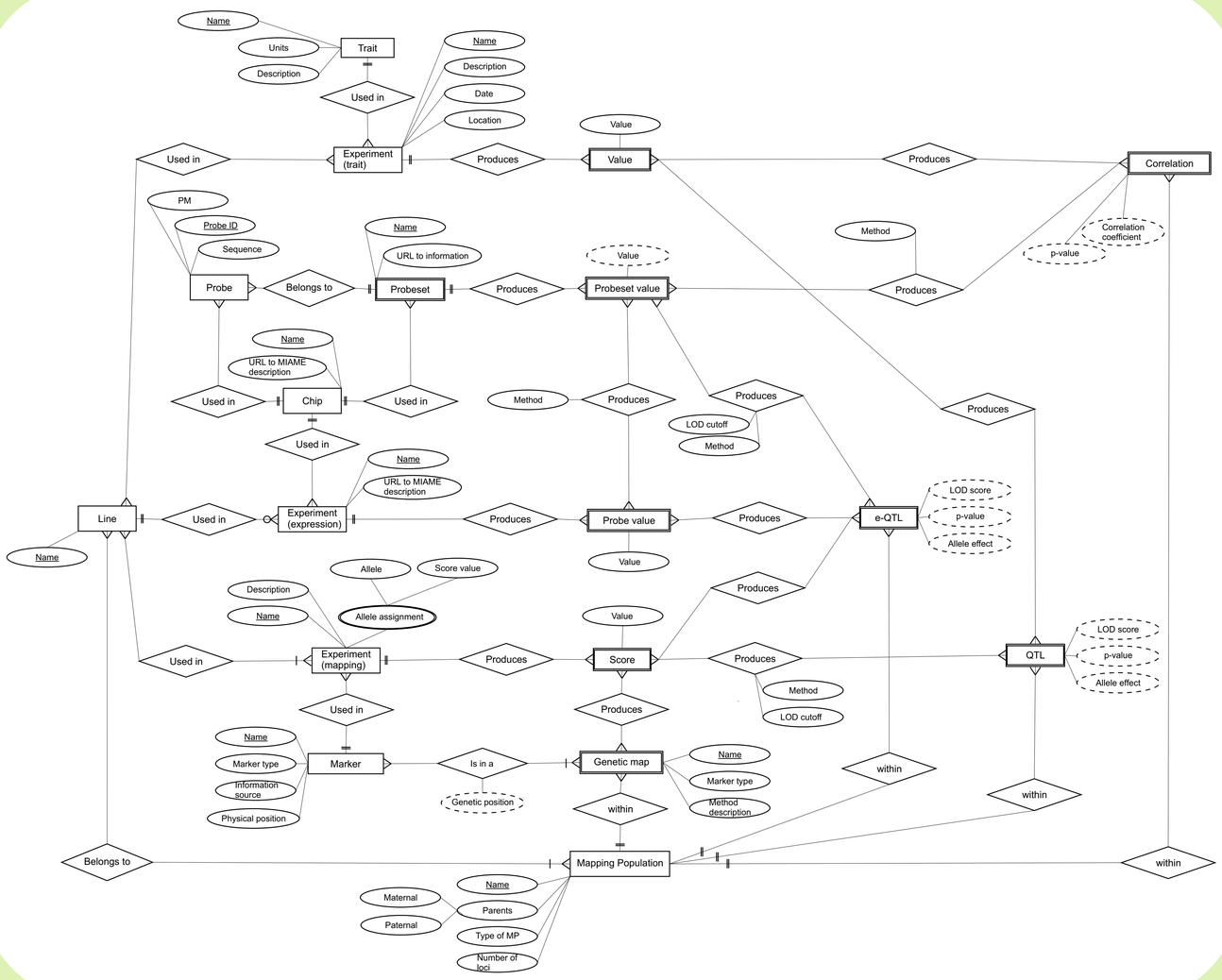
The practical application is aimed to facilitate the data exchange from the parallel, large-scale QTL mapping experiments. Quantitative traits (also known as polygenes) as opposed to mendelian traits (monogenes) are determined by many genes. Chromosomal region or regions harbouring genes that contribute significantly to the trait phenotype are identified by QTL mapping. Traditionally QTL mapping experiments deal with relatively few traits, no more than several dozen or less therefore the standards are quite loose. With the development of the large scale parallel analysis techniques it has become possible to capture and analyze many quantitative biochemical parameters very efficiently and reproducibly.

One of the most widely used (and arguably most informative) large scale parallel analysis technique is based on capturing mRNA (messenger ribonucleic acid) abundance. Modern expression profiling arrays are able to detect mRNA abundance of 20,000 to 50,000 genes simultaneously. Thus, in the context of the QTL mapping experiment, mRNA abundances of 20,000 to 50,000 genes are assessed whether they behave as a quantitative trait. The expression profiling data have a well established standard - MAGE-OM (Microarray and Gene Expression Object Model). To link these data with the QTL analysis results a different OM (object model) is required. To make it universally appealing and extend the possibilities of the applications the QTL OM should be applicable to the legacy data stored in a vide variety of sources.

XML Technology

XML (Extensible Markup Language) is a rapidly developing technology that has been very successfully applied to solve a range of problems as diverse as messaging for web services and exchange of medical record information. The XML technology allows effectively separate the data layer from the presentation layer of the application. If used correctly, this separation ensures that the data can be stored in one place only and therefore maintained in a consistent state. An appealing feature of XML is that with the current developments of parsers and XSLT the XML is “self sufficient”. The data can be stored, described and displayed using the same technology. However it needs to be noted that several areas of XSL are still in development stage.

A 2001 review of the XML technology used in bioinformatics stated that XML gives the means for defining strongly structured documents so computer programs can easily navigate through them and access relevant pieces of information. This sums up the most useful trait of the XML technology. Indeed the 2005 list of XML projects form OASIS (Organization for the Advancement of Structured Information Standards) lists a whole array of fields in life sciences where XML applications have spawned large scale projects.



Simplified ERD (Entity Relationship Diagram) of e-QTL data.

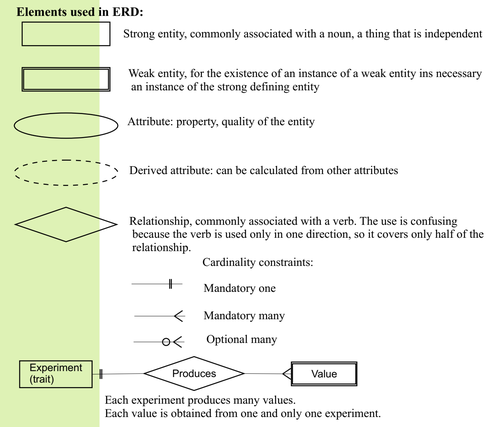
W3C XML Schema

W3C XML Schema is a W3C recommendation for the language describing the allowed contents of a class of XML documents. The XML Schema is one of the successors of Document Type Definition language (DTD). While DTD is still widely used to specify the contents of an XML document, it does not provide all the functionality desirable for the current success of the XML technology. The XML Schema widens the use of namespaces, user defined and built in data types, the differential control over document contents, inheritance and embedded documentation. However in comparison with DTD the XML Schema is much more complicated markup language to use.
The W3C XML Schema is a powerful development within the XML technology that offers significant advantages if used appropriately.
An XML Schema instance is an XML document that follows the Schema specifications and describes the contents of the XML documents validated with that schema. As an XML document XML Schema itself provides a sophisticated example of the power of the XML technology as well as offers all the advantages that an XML document has. These include the use of the XML validating parsers, the XSLT transformations, the universal use in applications.

Entity Relationship Diagrams

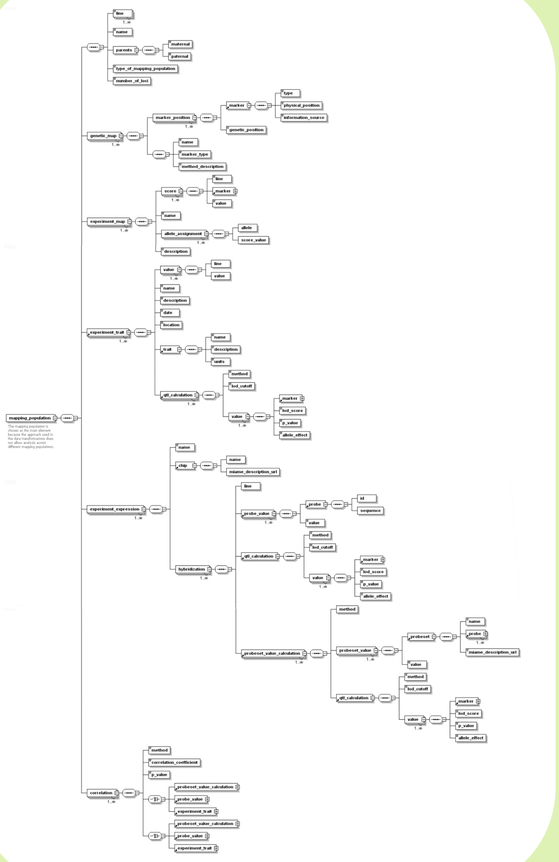
Entity Relationship Diagrams (ERD) and Enhanced Entity Relationship Diagrams (EERD) are used for formal visualisation of business rules that govern how the data are handled and stored. In modern computer science the use of ERDs and EERDs is largely superceded by UML, however the diagrams have the advantage of being comparatively simple and intuitive to use.

Elements used in ERD:

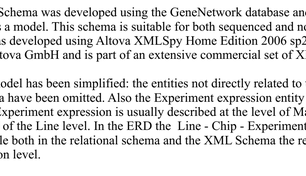


List of Abbreviations

DBMS Database Management System
ERD Entity Relationship Diagram
EERD Enhanced Entity Relationship Diagram
eQTL expression Quantitative Trait Locus
MAGE ML MicroArray and Gene Expression Markup Language
MAGE OM - MicroArray and Gene Expression Object Model
MAML - Microarray Markup Language
ML - Markup Language
mRNA messenger RiboNucleicAcid
OASIS - Organization for the Advancement of Structured Information Standards
OM Object Model
QTL Quantitative Trait Locus
RDBMS Relational Database Management System
SCRI Scottish Crop Research Institute
SVG Scalable Vector Graphics
UAD University of Abertay
UML Universal Modelling Language
W3C - The World Wide Web Consortium
XML Extensible Markup Language
XSLT Extensible Stylesheet Language Transformations



Symbols used in the XML Schema:



Discussion.

The e-QTL XML Schema was developed using the GeneNetwork database and SCRI Affymetrix Barley1 chip mapping experiment data as a model. This schema is suitable for both sequenced and not sequenced genomes. The XML Schema for the e-QTL data was developed using Altova XMLSpy Home Edition 2006 sp2. This software package is available as a free download from Altova GmbH and is part of an extensive commercial set of XML editing applications.

The e-QTL data model has been simplified: the entities not directly related to the e-QTL data and transformations necessary to obtain these data have been omitted. Also the Experiment expression entity has been simplified - the hybridizations or the Line - Chip - Experiment expression is usually described at the level of Mapping population - Chip Experiment expression instead of the Line level. In the ERD the Line - Chip - Experiment expression would describe a single hybridization, while both in the relational schema and the XML Schema the relationship is translated to the appropriate Mapping population level.

Also many entities that would be appropriate in an application (sequence data, gene data, sample description have either been omitted altogether or substituted by a URL to a different data source for the sake of clarity. At these points the data model becomes an add-on to the existing standard data models, like MAGE OM or Gene Ontologies.

The comparison between the relational database schema and the XML Schema suggests that biological data are much easier to translate to a hierarchical model than relational. Also the XML Schema visual display is more intuitive than the complex network of relationships between the tables in RDBMS.

ERD usually is a simple and effective tool to model the data. It may not be essential to use it to model trivial data or data with simple entity relationship structure. However in more complicated cases thorough studies of business rules and ERD is very beneficial. In the case of QTL and e-QTL careful studies GeneNetwork MySQL database structure as well as other databases (Gramene, Grain Genes) suggested that the data model itself is essential to develop optimal relational database schema with the right level of normalization for the application. Since the QTL and e-QTL data in particular have a complex and variable structure the ERD becomes also a useful tool in understanding the relationships between the different components of the model and choosing the right level of simplification for each particular application.

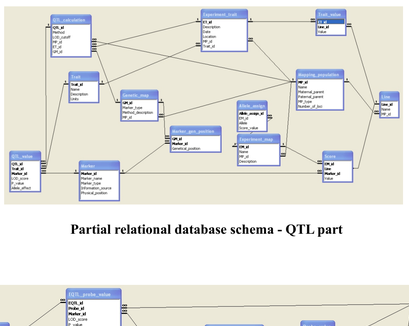
More extensive use of data modelling by either ERD or UML would lead to the use of better developed data models in biology applications, which in turn will facilitate more effective applications. Also it could aid the development of standard markup languages for biological data in mathematics (MathML) and chemistry (ChemML).

XML Schema itself is a valid and well-formed XML document therefore all the advantages of the XSLT transformations possible in an XML file can be achieved using the XML Schema itself. This means that the provided graphical image of the e-QTL schema can easily be transformed into subschemas of simpler structure which are more practical to use.
This is a serious advantage of the flexible XML technology over relational structures where the changes in schema are much harder to carry out.

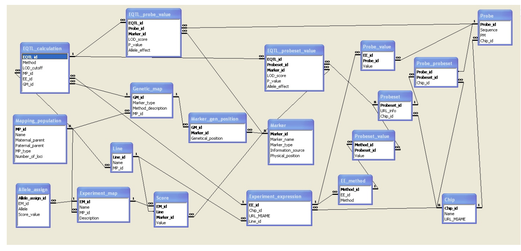
More information at http://barleygenome.net/xml

Acknowledgements:

Special thanks to Dr. Robert W. Williams, Hongquiang Li, Arturo Centeno and Zhaohui Sun from Department of Anatomy & Neurobiology, University of Tennessee Health Science Center, Memphis, TN, USA for their help and insight.



Partial relational database schema - QTL part



Partial relational database schema - e-QTL part

Both images represent partial schemas of a single database with 24 tables.
The table names are in blue rectangles on the top.
The primary keys are bold.
The number 1 above the relationship line denotes "the one" side of the relationship.
The symbol ~ above the relationship line denotes " the many" side of the relationship.

The relational schema for the e-QTL data model was developed as a side product of the XML Schema using Microsoft Access 2002. It is a simplified version of a e-QTL relational data model. The schema was developed for demonstration purpose only and is not intended to be used as a database. However it demonstrates the principles of developing the relational schema from ERD and the relationships between the tables are accurate. The database currently is in Normal Form 2 - the partial dependencies have been removed.

The complex relationship structure is due to the large number of "many to many" relationships in the data model.

Simplified relational database schema for e-QTL data.