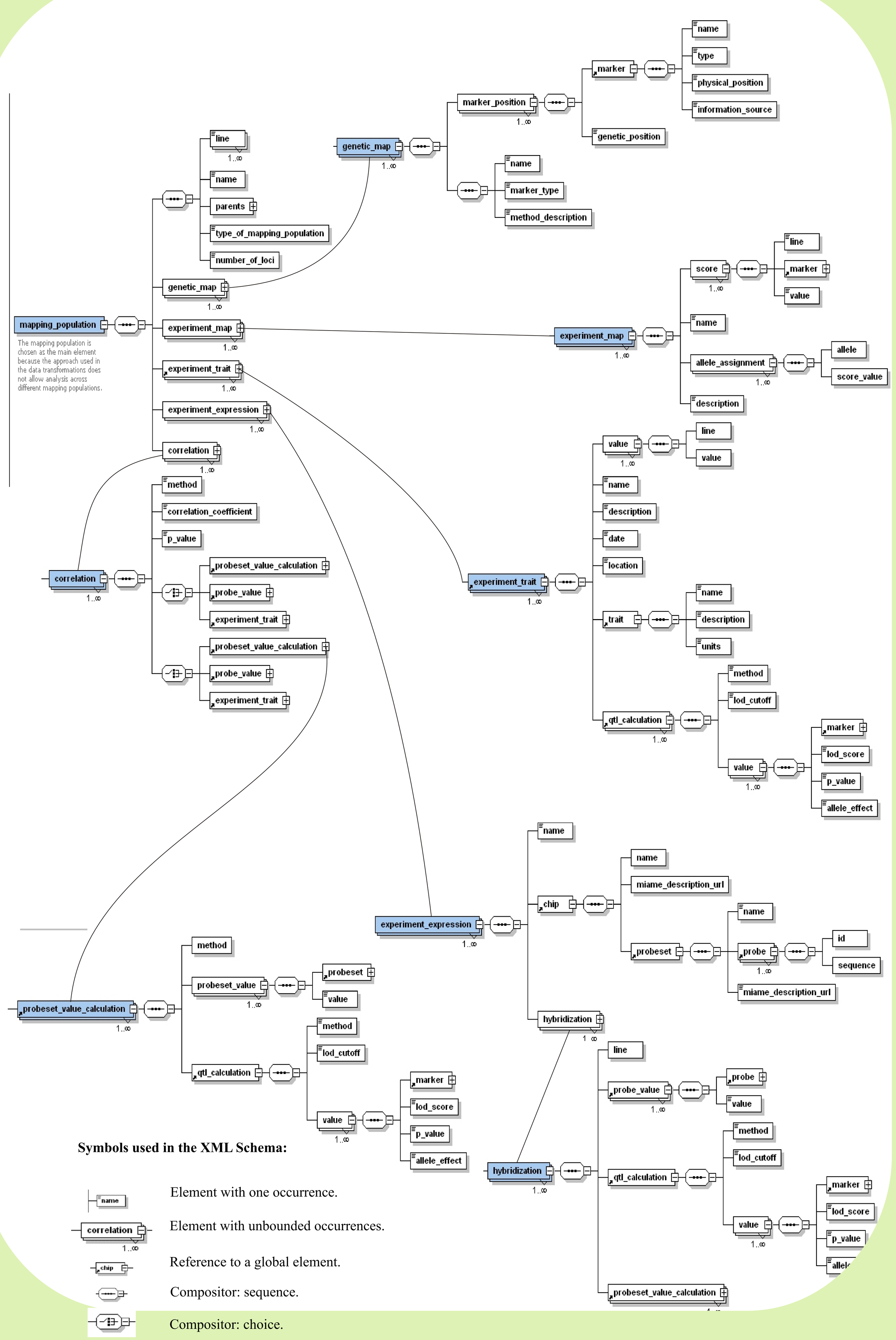
XML schema for the trait, genotype and mRNA expression data

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

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

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

Simplified XML Schema of e-QTL data.



Introduction

The use of XML (Extensible Markup Language) technology facilitates data exchange, legacy data incorporation and the development of lightweight, flexible, web-based graphical data display.

We developed an XML schema for the integration of the trait, genotype and large scale mRNA profiling data in accordance with the data structures exploited by MAGE-OM, GeneNetwork, GrainGenes, NCBI, Gramene and other acknowledged sources. The schema complies with the Second Edition of XML Schema Recommendation by W3C (World Wide Web Consortium).

Datasets from the trait QTL (Quantitative Trait Loci) and eQTL (expression Quantitative Trait Loci) mapping experiments in barley (*Hordeum vulgare L.*) were used as a model to develop corresponding XML 1.0 documents (http://barleygenome.net/xml) to test the schema.

Background

The eQTL - expression Quantitative Trait Locus

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 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.

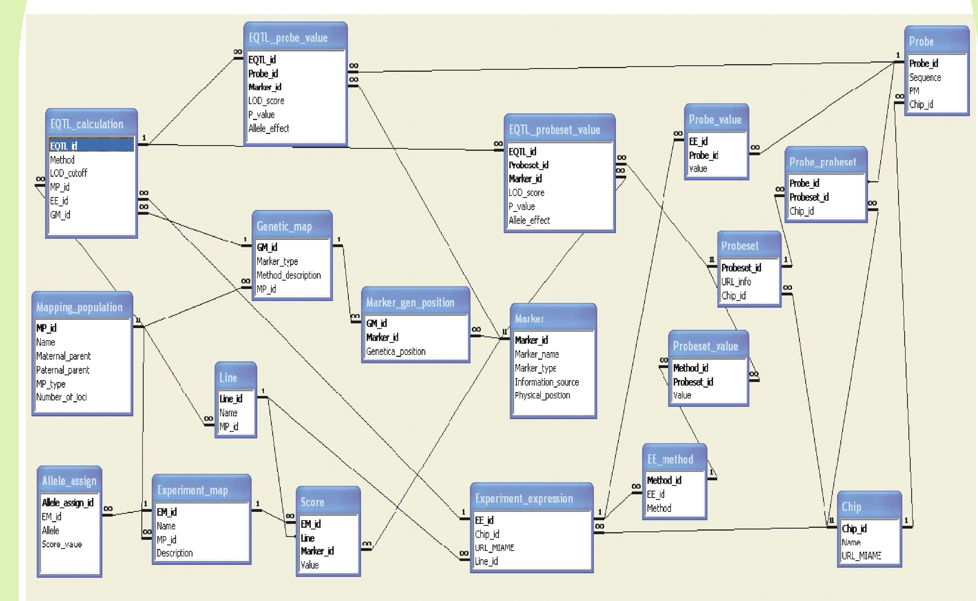
XML gives the means for defining strongly structured documents so computer programs can easily navigate through them and access relevant pieces of information. This sums up the most useful trait of the XML technology. Indeed the 2005 list of XML projects form OASIS (Organization for the Advancement of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

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. 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.

Simplified relational database schema for e-QTL data.



Partial relational database schema: e-QTL part (one of 24 tables).
Figure represents partial schema 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 part of the analysis of the XML Schema using Microsoft Access. 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.

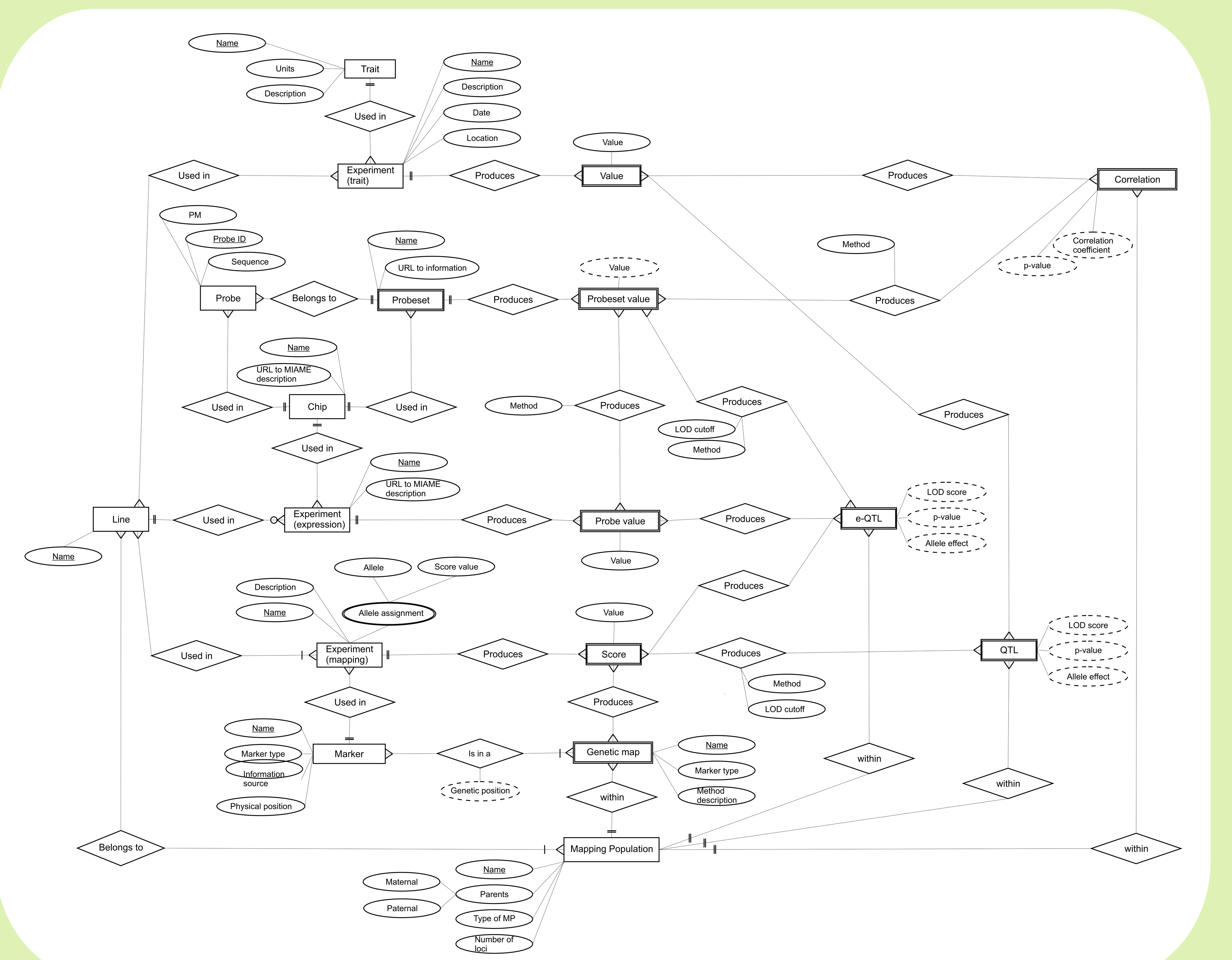
Drawbacks of the use of relational DBMS in biology

While the commercial and industrial data usually have quite clear and recognizable relationships between the entities (the customer places and order for an item, the order gets delivered, etc.) scientific data are different. During the design phase of the experiment the relationships between the entities are as clear as in commerce and industry. However with the rapid development of the technology and accumulation of vast amounts of data new questions and new relationships between the previously established entities as well as new entities emerge routinely. These features demand incorporating new data as well as data storage technology flexible enough to accommodate the changing needs. The relational DBMS are not designed to accommodate changing relationships especially if the schemas used in an application are designed with a specific task in mind deviating from the natural relationships between the entities. If this is the case, the database is ill suited to quick changes.

Excessive use of synonyms in naming instances of various entities or indeed the entities themselves complicates the relational schemas because each multivalued attribute is required to have a separate table to obtain a normalized database. Large number of tables increases the complexity of queries. The multivalued attributes are easier to handle within a hierarchical structure where they are rendered as multiple elements of the same kind.

The complex relationship structure is due to the large number of "many to many" relationships in the data model. The relational schema contains 24 tables with 102 fields altogether. These numbers are close to the 109 elements within 28 sequences in the XML Schema. This means that the semantic content of both models is similar and that the models are equally complex.

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



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. A well designed XML Schema can be a valuable aid in understanding the data structure especially in the cases when the data are complex. The development of the XML Schema for the eQTL data was based on the simplified EERD derived from an example data set and the workflow to obtain it.

The integral XML Schema of the eQTL data can be split into sub schemas describing the essential data sets used in eQTL analysis. Each of the XML Schemas for core entities can be used on their own as separate XML schemas for the corresponding data sets. The consequences of using different root elements were investigated and a conclusion made that a root element consistent with the workflow allows developing an XML Schema with fewer nesting levels but the same semantic content. The comparison between the XML Schema graphical view and the relational schema of a database developed from the same EERD showed that the hierarchical structure is easier to understand.

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.

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

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

Acknowledgements:
Special thanks to Prof. 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.

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.

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.

In the case of QTL and e-QTL careful studies GeneNetwork MySQL database structure as well as other databases (Gramene, Grain Genes) suggested that it 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 leads to the use of better developed data models in biology applications, which in turn facilitates more effective applications. Also it could aid the development of standard markup languages for biological data in mathematics (MathML) and chemistry (ChemML).

Elements used in ERD:

Strong entity: commonly associated with a noun

Weak entity: an instance depends on the existence of an instance of a strong entity

Attribute: property, quality of the entity

Derived attribute: can be calculated from other attributes

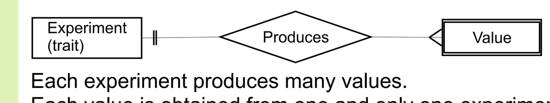
Relationship: commonly associated with a verb.

Cardinality constraints:

Mandatory one

Mandatory many

Optional many



Each experiment produces many values.
Each value is obtained from one and only one experiment.

The entities related to the array (Experiment(expression), Chip, Probe and Probeset) are also present in the MAGE OM. If necessary through these entities this schema can be linked to the corresponding classes of BioAssay, Array, Feature for both Probe and Probeset in MAGE OM.

In the EERD these points are denoted with the “URL to information” attribute. The naming conventions of the MAGE OM were ignored because in the attempt to generalize the descriptions of the arrays and features on the arrays the MAGE OM developers have devised a language very different from the one used in the laboratory to describe the same objects. All the entities used in this EERD can be described as subsets of the respective entities used in MAGE OM.