Workshop on Data Integration within the International *Arabidopsis* Community

Meeting Report

A workshop on data integration within the international *Arabidopsis* community, organized under the auspices of the Multinational Arabidopsis Steering Committee (MASC) and supported by the US National Science Foundation (NSF), took place at The Institute for Genomic Research (TIGR) on April 18-19, 2005. The purpose of the workshop was to respond to previous concerns, voiced by MASC, that better and more comprehensive data integration from both large and small data providers was required if the research community was to derive full benefit from the diverse resources and datasets being developed world-wide.

Aim of the workshop

The Multinational Coordinated *Arabidopsis thaliana* Functional Genomics Project Annual Report, 2004 makes several recommendations intended to further the goals of the current international *Arabidopsis* functional genomics projects. One of these is:
"Improved capabilities and integration of Arabidopsis database(s), with better ways to locate data and strongly enhanced mechanisms for import of data from individual researchers. In order to do simultaneous queries and analyses on large datasets, these finally have to be merged into one central database or a well-integrated network of databases."

The workshop brought together data providers and users to develop a strategy and propose solutions to address these issues. While some possibilities and experiences were presented, the main focus of the workshop was to articulate specific objectives and recommend strategies that could form the basis for discussion with the community, e.g. at the Arabidopsis conference June 2005 in Madison, and for efforts to implement these objectives. In addition to a roadmap of next steps, a number of self-identified working groups were set up to address more specific topics.

Attendees

The workshop was attended by representatives of MASC, wet lab-oriented biologists as users, data providers, database hosts such as TAIR, application developers, data integration projects such as BioMoby and PlaNet, and funding agencies including the NSF, the European Commission and the USDA. For a complete list of attendees please see the appendix.

Day 1: Requirements analysis

After a welcome by Chris Town (TIGR) and an overview of the workshop's goals by Heiko Schoof (MIPS), the first morning consisted of a series of short presentations by a spectrum of individuals intended to be representative of users and small and large data providers. This was followed by a number of presentations on current and possible future solutions for data management including TAIR and PlantGDB as well as solutions for distributed database interoperability (DAS, BioMoby, PlaNet).

In the course of this discussion, a list of data types currently being handled and generated was compiled. This showed that the complexity of the integration problem has been continually increasing in recent years, and while there are many encouraging experiences in integrating sequence and related data,

expression data proves more difficult. An even greater level of difficulty is expected for more qualitative data including phenotypic data, and protein interaction data.

The afternoon was started by a series of short discussion starters on the requirements for data integration from the viewpoints of computer science and database or application developers. Combined with input from the user and data generator perspectives, the following requirements were identified:

- The vision of a one-stop-shop, so that users need no longer search for and visit a number of different sites on the web and compile the information themselves, but can access all information from a single entry point. This need not mean that all information is physically located in a single database, or accessible through a single web site: Alternative user interfaces are of value to end users who can select an interface appropriate to their problem. However, the user should be able to navigate to all data from the entry point. Integration through web links was considered sufficient for this.
- Simple search interfaces.
- Multiple, domain- and problem-specific interfaces and views of the data.
- Integration of phylogenetic data and tools, such as orthologs/paralogs and phylogenetic trees, with the gene/genome data.
- Simplified update of information and capture of community knowledge.
- Availability of downloadable, integrated core datasets.
- More user-friendly, powerful and versatile visualization tools.
- Intuitive user interfaces developed according to human-computer-interface standards.
- The ability to build and use data analysis pipelines that connect several tools without the need to invoke each step individually.
- Current, detailed and comprehensive annotation, including knowledge collected from literature.
- High-quality data resources and long-term archives
- The ability to retrieve data from analysis and visualization tools, e.g. a selected set of proteins, and to input these into further tools.
- The ability to create and manage groups or sets of data.
- The ability to save and share work sessions.
- Standards for data formats (common syntax).
- Enforcement of low level analysis standards.
- Standard Operating Procedures, e.g. for GenBank submissions, to ensure that metadata is entered consistently (e.g. "sequence class" for mRNA-derived sequences).
- Standardized vocabularies, ontologies and terms, as well as defined usage of terms and tokens (common semantics, e.g. a common term for "gene name" and a common usage thereof only for the genetic name or only for the locus tag/identifier).
- Coordinated naming, ensuring uniqueness and lack of ambiguity, and the use of generally recognized and accepted names e.g. in publications.
- Manual curation of links between different sets data and of synonym lists.
- A common user management and login for multiple sites.
- Common interoperability standards and programmer interfaces to enable consolidation of development efforts.
- Public, machine-discernible interfaces to data and analysis resources.
- A data discovery service.
- An easily extensible and open architecture that at the same time could prevent confusion that might be caused by illicit data sources hosting nonsensical or wrong data.
- Synchronization of core data like the genome sequence/assembly between data providers.
- Provenance information, experiment descriptors and evidence metadata linked to all data.
- Documentation and tutorials.

Some of the questions that were raised and articulated during the discussions were:
How do we enable high-throughput data integration? While data can be integrated from multiple sources manually, this is not feasible given the large scale of current data sets.
How do we ensure availability of high-quality data? While enabling widespread availability of data what can or should be done to ensure that the data are reliable and of high quality?
To some of the workshop members, there was conceptual overlap between data integration and broadly based data mining. Hence there was quite a bit of discussion on data mining and the kinds of tools and data sets that the community might need.

To approach these requirements, service oriented architectures as implemented e.g. in Biomoby, PlaNet or Toolbus were juxtaposed to integrated data warehouses as implemented e.g. at TAIR. Discussion quickly made it clear that no single approach will answer all requirements. E.g., on the one hand, centralized curation of data can to some extent ensure quality, on the other hand, flexibility, extensibility, comprehensiveness and diversity can be more efficiently realized in a service-based, distributed system.

At several points, the discussion digressed into problems specific to a particular domain of data, which then served to illustrate fundamental prerequisites for comparability or integration of data: E.g. the necessity of controlled and carefully described experimental conditions for comparing microarray data from expression analysis, or the usefulness of a universal and unique identifier for microarray experiments were mentioned. This illustrates that data integration is not an isolated topic, but must be coordinated with the experimenters, large-scale data generators and users, as some prerequisites already need to be met before the data are generated. As a result, communication with relevant groups, initiated through the MASC subcommittees, was proposed.

Day two: Recommendations

The second day of the meeting began with a presentation from the RIKEN group on their efforts in data generation and provision and the discussion of some use cases. This was followed by a demonstration of the Taverna[1] Workbench, a web service client that allows the construction of workflows to retrieve and integrate data from a range of web service providers. This showed how database interoperability through web services can be used to generate analysis pipelines, integrate data from distributed sources on the fly and retrieve and store sets of data.
A presentation on semantic web technology demonstrated a technological perspective and a working prototype for machine-based data integration as implemented by the Semantic MOBY project. The increasing complexity of heterogeneous data and the necessity of large-scale manipulation of information and knowledge will make it necessary to enable machine-based integration that no longer requires human intervention.
Three break-out groups focused on specific problems:
1) Microarray expression data
2) Genome annotation and XML standards
3) Whitepaper draft
These returned with several action items that are included below.
Finally, the items for recommendations and outcomes of the workshop were discussed. There was general agreement on standardization efforts on controlled vocabularies, naming/identifiers, data formats and tokens. For data availability, both centralized, integrated data warehouse providers that are able to ensure quality through manual curation where necessary were seen as essential, as well as the implementation of a service-oriented architecture that allows easy addition of data sources and simplifies their utilization

[1] http://taverna.sf.net

both for retrieval of the data and for development of analyses and tools. Both approaches are in place and will continue to be developed. TAIR emphasized their ongoing dedication to integration, curation and provision of high quality data, and about half a dozen individuals indicated that they planned to initiate some form of web service from their own institution, augmenting the existing services.

Outcomes of the Workshop

Over the course of the workshop, the contents and functionalities of a wide spectrum of databases and data types were discussed quite extensively. There was general agreement that given both the magnitude and diversity of the current data sets and the anticipated growth, there is need to store and curate core essential data (e.g. annotations) into a centralized database (e.g. TAIR) but that it is unrealistic to try to collect all data, data types and services in a single or a few centralized locations. An approach that has proven successful in many other scenarios is a service-oriented architecture: While data integration is the ultimate goal, the first step is to achieve interoperability. This entails the definition of standards for data formats and semantics.
Although this workshop was restricted to integration of different data types and data sets pertaining to *Arabidopsis*, there was widespread agreement that whatever direction was taken to accomplish this goal, it should be readily applicable beyond *Arabidopsis* to embrace other species and coordinate with other communities.

Summary of Future Objectives:

To work towards more widespread and comprehensive data integration it is necessary to
-Ensure easy access to a comprehensive, integrated and high quality core data set
-Establish links between different data types
-Develop analysis pipelines and simplify access to them
-Facilitate generation of specialized (custom) datasets and combinatorial queries on data
-Develop and provide visualization and analysis tools

Recommendations:

To achieve the above objectives, the group makes the following recommendations

1. Create standardized tokens (e.g. <tags>) and subsequently ontologies to define the semantics of data representations in several realms:

- DNA sequence annotation,
- RNA properties
- protein properties
- metabolome
- phenotype
- evolutionary relationships

Examples: AGI Locus Code, Gene name/symbol, GO_term, expression experiment ID. This should be done in collaboration with data producers, data providers and analysts.

2. Propose standard data exchange formats for the above data types. If possible, use existing standards, and promote adoption of these standards by the community at large.

3. Support existing efforts to create controlled vocabularies/ontologies to describe e.g. plant anatomy (PO), phenotypes (PATO), biochemical function (in more depth than GO?) and localization (GO), sequence annotation (SO).

4. Create and promote the use of experiment and evidence descriptors to attach provenance information to the above data types for comparability and quality assessment purposes.

5. Form a working group comprised of both informatic and biological expertise (database developers, providers and users) to explore, assess and recommend technology for database interoperability.

6. Provide training and tutorials and documentation to both bioinformaticians and biologists to facilitate *Arabidopsis* data production, mining, analysis and integration.

Mechanisms

The following mechanisms were proposed to work towards these goals:

- Tutorial workshop at the *Arabidopsis* conference (Copenhaver, Schoof, Town)
- Tutorial documents on Web pages (Huala)
- White paper (This document, to be widely disseminated; Schoof and Town with input from all participants)
- Work groups on data types and formats to be developed in collaboration with the MASC subcommittees
- A database interoperability and data integration workshop at the end of 2006 that will bring together people actively involved in the efforts recommended by this workshop.

Action items

The following specific action items were identified to help maintain the enthusiasm and momentum generated by the workshop, to propagate its findings and to begin implementing the recommendations:

- Compile a comprehensive *Arabidopsis* mRNA expression dataset (Witt, TAIR, NASC)
- *Arabidopsis* conference workshop (Copenhaver)

- Initiate committees on (with MASC)
    - ontology development (Town)
    - semantic definitions and (Town)
    - data format standardization (Town)

- Report on existing efforts outside of MASC on
    - ontology development (Rhee)
    - semantic definitions and (Schoof)
    - data format standardization (Schoof)

- Build a searchable list of data providers and tools (Witt)

- Publicize the need for and work towards availability of all data in more global data providers such as Genbank/EMBL, GEO/ArrayExpress, BIODAS, BIND

- Conduct a use case survey (Copenhaver, Rhee, Gribskov)

- Prepare a note to a plant journal on the workshop and its outcomes (Town)

Appendix

Agenda

Monday, April 18

08.15 Shuttle from Quality Suites
08:30 arrive TIGR, coffee

- 09:00 Welcome (Chris Town)
- 09:10 Introduction of participants
- 09:20 Short introduction to aims and agenda of the workshop (Heiko Schoof)
- 09:30 Short presentations by representative stakeholders (database users and providers). What are their needs, interests and aims
    - Greg Copenhaver
    - Rodrigo Gutierrez
    - Sue Rhee: Overview of data handled at TAIR
    - Jen Sheen
    - Blake Meyers
    - Joe Ecker
    - Sean May
    - (Each of the above will make a 5-7 minute statement of their perspective as a catalyst for general discussion. This will be accompanied by the generation of an itemized list on e.g. a flipchart: What data are we talking about?)
- 10:30 break
- Past and present approaches and solutions:
    - 11:00 Strategies and solutions at TAIR (Eva Huala)
    - 11:10 Standard operating procedures and good practices at PlantGDB/AtGDB (Volker Brendel)
    - 11:20 DAS (Foo Cheung)
    - 11:30 BioMoby (Ben Good)
    - 11:50 PlaNet (Rebecca Ernst)
- 12:00 Q&A and discussion

12:30 Lunch

14:00 Round-table discussions to develop recommendations for data integration and availability (Chair: Heiko Schoof)

- 14:00 Basic concepts of data integration: The what, how, where or nuts and bolts (Michael Gribskov)
- 14:10 What are prerequisites for user-friendly display and application development? Experiences of VBI software developers. (Bruno Sobral)
- 14:20 What is needed: The prerequisites for, and future of, data integration. (Damian Gessler)
- 14:30 jointly sum up prerequisites and formulate these as objectives for the "recommendations". Key questions:
    - What is required on the data provider side?
    - What is required by data generators/experimenters?
    - What is required on the client/average user side?
- 15:00 How can those prerequisites be met? Open discussion. Key questions:
    - What know-how and existing technology is already there?
    - What needs to be centralized, where is distribution necessary?
    - What profits from warehousing, where is federation profitable?

    - What standards and agreements are required?
- 15.30 Short Break – refreshments, leg-stretch etc.
- 15.45 Continue discussions
- Summarize by a list of solutions matched against objectives they target.
- ~17.00 End of session
- Reception and dinner at TIGR

Tuesday April 19

- 08.00 and 08.15 Shuttle from Quality Suites
- 08:30 arrive TIGR, coffee
- 09:00 Summary of the previous day
- 09:15 Motoato Seki – Data Resources at RIKEN
- 09:30 Heiko Schoof – Demonstration of Taverna Workbench for BioMoby
- 10:15 Damian Gessler – Semantic Moby
- 10:30 Break
- 11:00 Round table discussion, selecting and discarding solutions proposed the day before:
    - What technological solutions best answer the tasks?
    - How can they be combined?
    - What standardization efforts are required?
    - What is practical and achievable in the short/long term?

12:30 Lunch

- 13:30 General discussion. Identification of break-out groups.
- 14:00 Break-out groups meet; begin drafting of outcomes and objectives.
- 15:00 Breakout groups reconvene; general discussion of content and wording of draft outcomes and objectives

Attendees

Doug	Becker	TAIR	dhbecker@gmail.com
Indridi	Benediktsson	EU Reseach Commision	Indridi.Benediktsson@cec.eu.int
Volker	Brendel	Iowa State University	vbrendel@iastate.edu
Foo	Cheung	TIGR	fcheung@tigr.org
Parag	Chitnis	NSF	pchitnis@nsf.gov
Greg	Copenhaver	University of North Carolina	gcopenhaver@bio.unc.edu
Gloria	Coruzzi	New York University	gloria.coruzzi@nyu.edu
Joe	Ecker	Salk Institute	ecker@salk.edu
Rebecca	Ernst	MIPS	rebecca.ernst@gsf.de
Damian	Gessler	NCGR	ddg@ncgr.org
Ben	Good	University of British Columbia	bmg@sfu.ca
Michael	Gribskov	Purdue University	gribskov@purdue.edu
Erich	Grotewold	Ohio State University	grotewold.1@osu.edu
Rodrigo	Gutierrez	New York University	rg98@nyu.edu
Eva	Huala	TAIR	huala@acoma.Stanford.EDU
Ed	Kaliekau	USDA-CREES	ekaleikau@csreeusda.gov
Sean	May	NASC	sean@arabidopsis.info
Blake	Meyers	Delaware Biotechnology Institute	meyers@dbi.udel.edu
Sue	Rhee	TAIR	rhee@acoma.Stanford.EDU
Heiko	Schoof	MIPS	h.schoof@gsf.de
Motoaki	Seki	RIKEN	mseki@gsc.riken.jp
Jen	Sheen	MGH, Harvard	sheen@molbio.mgh.harvard.edu
Ian	Small	INRA	small@evry.inra.fr
Bruno	Sobral	VBI	sobral@vbi.vt.edu
Tatiana	Tatusov	NCBI	tatiana@ncbi.nlm.nih.gov
Sakurai	Tetsuya	RIKEN	stetsuya@gsc.riken.go.jp
Chris	Town	TIGR	cdtown@tigr.org
Yves	van de Peer	VIB-Ghent	yves.vandepeer@psb.ugent.be
Bernd	Weisshaar	University of Bielfeld	weisshaa@cebitec.uni-bielefeld.de
Jennifer	Weller	George Mason University	jweller@gmu.edu
Isabell	Witt	MASC	isawi@duke.edu
Manfred	Zorn	NSF	mzorn@nsf.gov