

# **Australian Winter Cereals Pre-Breeding Alliance**

## **Bioinformatics Review**

**July 2006**

**Prof Rudi Appels (Chairman)**, MPBCRC, Murdoch University/Dept of Agriculture and Food WA

**Drs Bill Taylor and Gavin Kennedy**, CSIRO, Plant Industry, Canberra

**Dr David Butler**, QDPI Toowoomba, Queensland

**Dr David Edwards**, LaTrobe University/Plant Science Centre, Victoria

**Ms Clare Johnson**, VAWCRC, Sydney, NSW

**Prof Diane Mather**, MPBCRC, University of Adelaide, SA

**Prof Richard Oliver**, Murdoch University, WA and GRDC Western Panel

## GRDC Bioinformatics review

### Conclusions/recommendations listed against terms of reference:

- Examine current systems that are being used to manage and utilise genomic and genetic information for wheat and barley improvement  
*Appendix 1 provides a summary*
- Determine what deficiencies exist, for example in capacity or lack of interfaces between different data management systems  
*Recommendation (#1) is that to facilitate the ease of data exchange between the different operational systems used among breeding programs, a common standard for data exchange (for example Extensible Markup Language, XML) needs be agreed upon in coordination with other ontology consortiums at an international level. Web-based systems are generally accepted as the mechanism for interfacing across the different bioinformatics systems.*
- Identify opportunities; discuss whether change is possible and desirable  
*Recommendation #2 is to focus on a service component in this area of activity. A key observation was that gaps in communication between different groups in Australia exist as well as gaps between the bioinformatics scientists and breeders. This observation was combined with the discussions around the fact that the databases require significant curation activity to provide quality assurance and adequate cross-referencing of datasets. This would include a contribution to the international effort to establish a phenotype-ontology (a common language) for breeding traits and could leverage work on Excel/XML data submission templates to assist metadata management already underway in ICIS and elsewhere (see Appendix 1(5)).*
- Make recommendations for national coordination and management, to be referred back to the Pre-breeding Steering Group  
*Recommendation (#3) is that models such as the GRDC funded National Statistics group and the Map Curation group (utilizing CMap) should be explored for the national coordination and management of bioinformatics related to breeding in Australia. Two levels of interaction were identified  
(1) workshops to engage users of the tools and to stimulate the development of software to solve specific issues of data handling.  
(2) workshops to bring together technically competent individuals to facilitate the implementation of a common standard for data exchange based and deal with the fundamental problems of definitions and ontology*

Recommendation #1 did not cover the opinion of Professor R. Oliver that all pre-breeding bioinformatics should be covered by adopting a single “system” – this position was specifically discussed by the group and it was not possible to reach a position that came close to this opinion.

Recommendations in an OECD report on bioinformatics and a paper describing XML are provided (attached).

Process:

- Several email exchanges to identify capacity around Australia and identify some key issues
- Phone conference Thursday June 15<sup>th</sup>, 2006
- Feedback on a draft document was received from all the participants in the phone conference discussions by Friday July 7<sup>th</sup> and the comments were incorporated into the final document.

Members of the discussion group:

- Prof Rudi Appels (Chairman), MPBCRC, Murdoch University/Dept of Agriculture and Food WA
- Drs Bill Taylor and Gavin Kennedy, CSIRO, Plant Industry, Canberra
- Dr David Butler, QDPI Toowoomba, Queensland
- Dr David Edwards, LaTrobe University/Plant Science Centre, Victoria
- Ms Clare Johnson, VAWCRC, Sydney, NSW
- Prof Diane Mather, MPBCRC, University of Adelaide, SA
- Prof Richard Oliver, Murdoch University, WA and GRDC Western Panel

Prof Vladimir Brusic (ACPF, SA/QLD) was invited but did not contribute.

Summary of capacity around Australia

Appendix 1 provides a summary of capacity in the bioinformatics area associated with breeding programs. Although not exhaustive it emphasizes that all the centres around Australia are well equipped with respect to hardware with generally complementary expertise – some overlap occurs and the discussion group agreed that this was generally a healthy feature of this area of research in Australia as it stimulates innovation.

Issues discussed

The application of bioinformatics, with developments led by end users to resolve real research bottlenecks, is now a common feature of data processing across groups in Australia. The resulting systems aim to allow users to search and link from field trait data through molecular markers and genetic maps, to sequenced genes, genomes and gene expression information, using a variety of data formats with which they are familiar. In addition, specific data processing tools are generally open source or developed locally for specific tasks. The different systems across Australia are generally supported by tools for data analysis and a security system which enables restriction and sharing of data on a user by user basis. Although some overlap occurs between different research groups, the discussion group agreed that this was generally a healthy feature of this area of research in Australia as it stimulates innovation.

A valuable part of the discussions was to hear about the experience of the network in place between the QDPI&F and NSW DPI through Katmandoo and the network of statisticians in NSW, SA and WA through the National Statistics Project led by Dr Brian Cullis. Web-based systems are generally accepted as the mechanism for interfacing across the different bioinformatics systems in place. It was the experience of these networks that the overlap among the various bioinformatics initiatives was not considered a major issue and that the different groups were more interested in the opportunity to communicate and share technology. All the participants in the discussion emphasized the need for user support and inter-operability between bioinformatics systems in place across Australia. It was noted that gaps do exist between the bioinformatics scientists and breeders and this also needed attention.

A key issue identified regarding the inter-operability issue, was the ease of data exchange between the different operational systems used among breeding programs. The data exchange can occur “table-to-table” on a large scale or via .xls or .csv files on a smaller scale, and there is uniform agreement that a single mega-system covering all the breeding programs is not required. Excel is widely preferred by laboratory staff for its tabular approach and ease of use, particularly as an import/export mechanism to the formats required by the proprietary packages used. These include MapManager, JoinMap, Q-Gene and MapChart which all accept flat text file input of similar (but not identical) format. In the case of ACCESS databases (common in the research community) data from this system is easily exported into a master database accessible through a web-based system with functionalities that can change as the community using the data evolves. The significant conclusion from the discussion on the ease of data exchange between the different operational systems used among breeding programs was that there needs to be an adoption of a common standard for data exchange based on Extensible Markup Language (XML). XML is a simple, very flexible text format derived from SGML (ISO 8879). It was originally designed to meet the challenges of large-scale electronic publishing, and is now playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere. One of the attachments provides a description of XML.

It was evident from the discussions that the databases require significant curation activity to provide quality assurance and adequate cross-referencing of datasets. In addition, across-program datasets need to take into account privacy requirements and therefore establish a strong security model which not only protects data from external viewing, but protects the existing data from inadvertent corruption/overwriting. The quality assurance and adequate cross-referencing of datasets (including an inventory of research projects and bioinformatics capabilities) has a significant service component and it was noted that this was labour intensive. Interest was also generated in developing an Australian contribution to the international effort to establish a phenotype-ontology (a common language) for breeding traits

## **Appendix 1**

### **Capability overview for bioinformatics activity in groups associated with breeding programs.**

#### **(1) WA Centre of Excellence for Comparative Genomics (Murdoch university/Western Australia) Matthew Bellgard**

Dedicated server room linked to the outside via two fire-walled network links. One 100Mb link to the Murdoch University Network and a dedicated 1Gb fiber link to IVEC. CCG internal network is 1Gb. Standard IT infrastructure includes; multiple web servers, LDAP server, firewalls, mail server, multi terabyte capacity (tape) backup server providing daily backups, relational database server (Dual opteron, 8GB RAM). Dedicated bioinformatics platforms include a Research machine (4-way opteron 16GB RAM) available to all staff for developing and performing analysis, a Compute cluster (16 x dual opteron 4GB RAM 300GB SCSI disk) and a network file server for running over 40 bioinformatics related tools (Blastall, clustalw, repeatmasker and so on), and a development server (Dual opteron 4GB RAM). Main public databases are stored locally for searching - the local copy of Genbank is updated weekly. Numerous custom datasets have also been created and can be searched using the compute cluster.

Interaction with WA breeding programs is via a direct collaboration with the Department of Agriculture and Food WA to upgrade PBGenesis and via the Molecular Plant Breeding CRC to establish specific software solutions to issues raised by researchers. The GRDC funded project on Map Curation is led from the Murdoch University/ Department of Agriculture and Food WA group

A key issue identified is the ease of data exchange, as required, between the different operational systems used among breeding programs. This can be solved by suitable software to allow table-to-table exchange on a large scale or via .xls or .csv files on a smaller scale, and there is uniform agreement that a single mega-system covering all the breeding programs is not required.

## **(2) DPI Victoria capabilities and requirements overview in Bioinformatics**

### **Dave Edwards**

The BASC bioinformatics hardware consists of an IBM 140 CPU Linux compute farm (70 nodes, each with dual 2.8GHz Xeon processors, 4GB RAM and 2 x 146GB SCSI hard drives). This is integrated with 1x IBM pSeries 615 and 2x IBM pSeries 655 (a total of 17 power 4+ processors and 66GB RAM), and the complete system shares 1.7 TB SAN disc storage, automated tape backup and appropriate connectivity. Additional hardware includes two sun solaris machines and several additional Linux machines. The majority of the system resides behind a secure firewall with access through the secure DPI network or externally using a VPN client. Four dedicated servers reside within a secure DMZ with both open public and password/certificate secured access for non-DPI based collaborators and public tools hosted by DPI.

### BASC bioinformatics modules

- The ESTDB module consists of a processing pipeline and database system for the processing, annotation, storage, searching and integration of expressed gene sequence information (ESTs). Molecular genetic markers derived from these sequences link to the MarkerQTL module. Related gene expression information is maintained within the microarray database.
- The MarkerQTL module is the most complex and houses information on molecular genetic markers; individuals, populations and associated phenotypic data; genotypes; traits; genetic maps; QTLs and genetic diversity data. This module is fully searchable and links to the CMAP, ESTDB and EnsEMBL modules. Custom data import and export ensures integration with breeders tools such as Agrobase I and II, MS Access as well as a variety of spreadsheet and external database designs, with new import/export scripts developed as required.
- The Microarray module houses gene expression information, predominantly produced using microarray technology, though other forms of gene expression data (eg. MPSS, SAGE) are compatible. Information includes experimental conditions, RNA preparation details, side information, gene expression values including raw and analysed data, links to original data files and links to the ESTDB and EnsEMBL modules. The Microarray module is supported by staff with specific training in this area and with access to a variety of specialist software.
- The EnsEMBL system was developed by EBI for the analysis of the human genome. This system has now been adopted by crop research communities. Within BASC, DPI Victoria have integrated EnsEMBL genomes for rice (initially developed as the Gramene database in the US) and Arabidopsis (developed by the University of Nottingham in collaboration with DPI Victoria). Within BASC, DPI Victoria has integrated proprietary data as additional layers, with pages created dynamically depending on the user rights and with full text searching and links to ESTDB, CMAP and MarkerQTL.
- The CMAP system developed by GMOD is becoming the accepted standard for viewing, comparing and integrating genetic map and QTL information. DPI Victoria have integrated public crop genetic map information with proprietary information

using this system and established novel systems for integrating related maps into consensus genetic maps.

- The ICIS system was developed by IRRI to enable breeders to load and store information pertaining to genealogy and phenotypic information. DPI Victoria is collaborating with IRRI to extend ICIS capabilities and integrate ICIS with the BASC system.
- The Shopping Cart Module is a secure data storage area for each user for data import/export to modules and PISE based programs.
- Molecular marker discovery tools – developed in house
  - SNPServer: Public web site hosted system for SNP discovery in any species.
  - AutoSNP: Stand alone system for SNP discovery in any species. Distributed under free academic licence.
  - SSRPrimer: Public web site hosted system for SSR discovery in any species. Software also distributed under free academic licence.
  - SSR Taxonomy tree: Public web site hosted system for SSR discovery in any species.
  - SSRPoly: Stand alone system for SSR discovery in any species. Under development. When completed it will be distributed under free academic licence and web enabled.
  - Sequence quality assessment: Tool to provide feedback on sequence trace quality – used in house, currently developing a web interface
  - Sequence annotation pipeline: Tool to create a portable nest of web pages of sequence annotation from a sequence input file – used in house, currently developing a web interface for sequence input
  - Microarray data analysis tools: several developments with academic/industry partners
  - Metabolite data analysis tools: several developments with academic/industry partners
  - PISE tools: these are web enabled tools, including all EMBOSS applications running on the core BASC cluster, with either email, shopping cart or web browser delivery of results. This currently houses over 150 bioinformatics tools. PISE wrappers are developed to provide access to new tools as requested by users.

### **(3) QDPI&F Plant Science Group**

#### **David Butler**

The Unit is currently staffed by two full time software engineers, shares the resources of a data management technician and has access to several biometricians attached to northern region germplasm development programs. The primary development environment is MS Windows based machines; however, the Unit has access to UNIX and Linux systems as required. Database systems are deployed in either MS-Access, SQLServer or MSDE for relational models or hybrid methods using array based storage for large applications such as DArT marker systems

#### Projects/Applications

- The Bioinformatics Unit is primarily focused on the needs of the region's applied germplasm enhancement programs.
- PBMASS (Pedigree Based Marker Assisted Selection System) has been developed as a desktop tool to assist plant breeders in utilizing pedigree and molecular marker when selecting breeding lines.
- The application provides graphical representations of pedigrees, linkage maps (colour coded for identity by descent (IBD) or identity by state (IBS)) and summarised phenotypic data along with the ability to:
  - Parse and generate Purdy (Purdy:1968) style pedigree strings
  - Manage aliases
  - IBS v's IBD calculations for selected genotypes / alleles
  - Store multiple linkage maps and molecular marker data for mapping and more general population structures.
  - Infer missing marker data
- Katmandoo is a joint project between NSW DPI and QLD DPI&F to develop a crop independent information system. The goal is to provide a robust platform for the implementation of advanced statistical, breeding and visualisation methods that integrate phenotypic data with known genetic information such as pedigree and molecular data, including high throughput systems. Katmandoo will incorporate the features of PBMASS in addition to the operational systems below.

#### Operational systems

We are currently developing or maintaining various operational systems to capture field and laboratory data, manage seed inventories and automate parental selection and crossing.

These are currently independent (for development and testing purposes) relational databases and associated applications that will be integrated within Katmandoo.

#### Networks

The Unit has strong links to NSW DPI through Katmandoo and the network of statisticians in NSW, SA and WA through the National Statistics Project led by Dr Brian Cullis.

**(4) Statement of Plant Industry's Bioinformatics Capability for GRDC  
Gavin Kennedy, Bioinformatics Leader, CSIRO Plant Industry, 31/05/06**

CSIRO Plant Industry manages the CSIRO Bioinformatics Facility. This facility consists of a 142 processor computer cluster, a 1.5 terabyte file server, database servers and web/interface servers networked via a 1 Gbit Ethernet. This facility supports our in house databases as well as mirror images of public databases such as NCBI Genbank, Swissprot, Trembl and TIGR sequence databases for Arabidopsis, Rice, Wheat and Grape. The facility is primarily used for high throughput annotation, assembly, alignment and genome mapping. Our in-house databases, (developed with CSIRO Mathematical and Information Sciences) include; the RGMIMS database that manages high throughput mutagenesis, crossing and phenotyping data, the GENA database to manage cDNA microarray data and the CEREALinfo database supporting sequence comparisons against the wheat and maize genomes.

Bioinformatics Support

The Plant Industry bioinformatics group assists CSIRO researchers in sequencing, sequence analysis, gene expression analysis, QTL analysis, statistics, data management, data visualisation and data publishing.

Bioinformatics Research

We also have research projects in scientific databases and data representation, regulatory element prediction. With CSIRO Mathematical and Information Sciences we have research projects in experimental design for multi-phase microarray experiments, small RNA prediction and modelling, massively multivariate data analysis.

GRDC Based Interactions

The Plant Industry Bioinformatics Group is not particularly concerned that there may be redundancy amongst the various bioinformatics initiatives. What we are interested in is the opportunity to communicate and share technology with these groups as well as determine how best to share data resources that will be useful to the researchers. Currently we would benefit most from agreements to share molecular marker information and practices to facilitate this data sharing. We would also like to learn from other groups how they approach metabolomics and proteomics data analysis, which are areas where we lack skills.

## **(5) ICIS Development Workshop 2006 and the CAGE implementation of ICIS**

### **Clare Johnson, Wheat CRC; GRDC CAGE Coordination Project**

UQ holds the GRDC grant for management of the CIMMYT-Australian germplasm evaluation (CAGE) data <[www.wheat-research.com.au/CAGE\\_index.html](http://www.wheat-research.com.au/CAGE_index.html)>. They are using an ICIS format, expanding on GWIS, which already holds historic Australian wheat pedigrees and reports on trials for numerous varieties: <<http://mendel.lafs.uq.edu.au:8080/ICIS5/>>. At the ICIS Development Workshop, 15–19 May 2006 at El Batan, Mexico, ICIS partners from CIMMYT, Nunhems, ICARDA, UQ, CIP, AAFC, University of Agricultural Sciences Bangalore and IRRI reported on progress to existing ICIS modules and formed work groups to coordinate future work.

Existing (open-source) ICIS functionalities include:

- Genealogy Management
- User Access Management
- List Management
- Data Management
- Inventory Management
- Gene Management
- Genetic Resources Information Management
- Location Management

Notable developments were the molecular marker module, system integrity and data validation tools from IRRI, high function multidimensional / OLAP developments including DataMart - Mondrian and DIVA\_GIS integration with ICIS, and CIPPEX, a molecular LIMS, by Edwin Rojas at CIP (Peru), and the Data Comparison Tool, a very user-friendly web interface front end for breeders to set up drag-and-drop multi-variable queries without the need for training in ICIS, developed by Shawn Yates at the Semiarid Prairie Agricultural Research Centre at Swift Current, Saskatchewan. Data is input via Excel templates designed to allow direct uploading of data to the ICIS database back end.

The Australian web-implementation of ICIS for the CAGE suite, to be developed incrementally over the next two years in collaboration with CIMMYT's Crop Research Informatics Laboratory (CRIL) and IRRI, will integrate the upgraded modules developed over the last year to make use of these functionalities, with differential security to separate public and private data. This will include the capacity to handle molecular marker data and DArT profiles, and the development of queries, visualisation and analytical techniques using CRIL's expertise.

The Developer group was interested in integrating the Gramene-derived ontologies feature from RGMIMS reported by Leakha Henry, CSIRO, into the ICIS suite.

A fuller summary of the ICIS Developers' Workshop 2006, including many of the presentations, is on ICISWiki at:

<[http://cropwiki.irri.org/icis/index.php/ICIS\\_Workshop\\_2006](http://cropwiki.irri.org/icis/index.php/ICIS_Workshop_2006)>

## **Appendix 2**

See attachments for a short manuscript on XML and an OECD report on bioinformatics