The URGI bioinformatic platform: an original information system to bridge genetic and genomic plant and fungal data

Delphine Steinbach[1], Joelle Amselem[1], Michael Alaux[1], Françoise Alfama-Depauw[1], Baptiste Brault[1], Nathalie Choisne[1], Victoria Domingez-Del Angel[1], Sophie Durand[1] , Benoit Hiselberger[1], Olivier Inizan[1], Véronique Jamilloux[1], Aminah-Olivia Keliet[1], Erik Kimmel[1],Nicolas Lapalu[1], Isabelle Luyten[1], Nacer Mohellibi[1] , Cyril Pommier[1], Daphné Verdelet[1], Sébastien Reboux[1], Marc-Henri Lebrun[2], Hadi Quesneville*[1]

[1]URGI (Unité de Recherche Génomique-Info), UR INRA 1164,
Centre INRA de Versailles, bâtiment 18, RD 10, 78000, Versailles, France

[2]BIOGER-CPP, UMR1290
Site de Grignon
av Lucien Brétignières, 78 850 Thiverval Grignon, France

*To whom correspondence should be sent: hadi.quesneville@versailles.inra.fr

Introduction

Efforts in genomics since the late 90s have led to major advances in understanding the biological process underlying the variation of traits, including key traits of agronomical interest. But a gap still remains for their application to the development of new cultivars beneficial to the society. New efforts are needed to bridge genomics and crop genetic diversity analysis. New programs should include extensive characterisation of genetic diversity in relationship with (i) new phenotyping approaches including proteomics and transcriptomics, (ii) mapping information at the genetic, physical and sequence levels. Ultimately, all data should be combined using appropriate methodologies and software to extract new knowledge.

The URGI platform aims at providing tools for such an approach. It allows biologists and bioinformaticians to store and retrieve data from a large number of origin. These data can be combined to extract new knowledge. The URGI platform can help biologists in the management and the valorisation of their data in connection with those obtained by the scientific community. The URGI platform also develops new tools to identify and to manage these connections. We have developed an information system called GnpIS and associated pipelines providing a complete annotation system.

GnpIS Architecture

GnpIS (http://urgi.versailles.inra.fr/gnpis) relies on cutting edge databases and data warehouse technologies. It is accessible *via* a public web interface, allowing high throughput storage and deep navigation in plant and bio-aggressors genomic and genetic data. GnpIS is a web based system composed of several relational databases. Their key feature is that they are connected through (i) sharing common tables and (ii) cross-reference links. This design strengthens data consistency throughout the system. Users can submit their data through an automatic submission portal offering data format validation facilities. Different data types can be uploaded: annotations, gene expression, proteomic data, DNA polymorphism, genetic and physical maps, genetic resources collections and phenotypes. Once inserted in the system, users can intuitively navigate through the data *via* the existing interoperability between our databases modules.

We consider each database with its web interface as a module of GnpIS. These modules are:

- Aster is the key and central module of the information system. It allows the interoperability between the data from different modules.
- GnpSeq allows to store and to query ESTs by clone name, library name, sequence, on (home-made) contig version or by external database identifiers.
- GnpMap allows the biologist (i) to browse genetic maps for several species in the same database, (ii) to query by markers, maps, trait, QTL, (iii) to display and compare several genetic maps (MapComparator) or physical maps (WebFPC, GBrowse, Cmap softwares). Links exists between loci and EST sequences data from GnpSeq to allow co-localisation between markers, genes and QTLs.

- GnpArray stores raw and normalized data as well as gene lists results from different kinds of expression experiments (microarray (cdna, probes), macroarray (high density filters), or technology based data: (affymetrix or nimblegen). Links exists between GnpGenome database for sequences mapped on the genome.
- GnpSNP allows users to query on the polymorphisms themselves (SNP, InDel, SSR), on genotypes, genes and alleles and to view them in a graphical way on genomic sequences in GnpGenome.
- GnpGenome database is based on GMOD Gbrowse and chado database model. It contains genomic sequence data and annotations. It allows an integrated and synthetic view of all the data of GnpIS mapped on the genome. It has links with other modules available through popup menu.
- Siregal stores passport data descriptors for plant genetic resource collections: the taxonomy, the country of origin, the pedigree and some phenotypes.
- Ephesis module (still in development) is dedicated to support the study of phenotypes through the genotype x environment (climate, soil) interaction. It will interact with Siregal and GnpSnp to investigate the relationship between genotype and phenotype.

Sophisticated query interfaces allow the users to easily (i) query all the data using a “google-like” approach or (ii) to cross informations with the Biomart system.

- A quick search tool based on Lucene technology, allows to search as a “google search” into all the databases of GnpIS by using precomputed indexes.
- An advanced search tool based on Biomart query management system. It uses database marts and datasets designed for quick access and by the use of fields combinations. Results are displayed in one unique page with clickable item lists, which serves as entry point to the GnpIS databases or as tablesheet data to export for further analysis.

An annotation system

Three annotation pipelines were developed:

- A structural annotation pipeline, based on *ab initio* and similarity gene finding softwares and the EuGene program to integrate all sources of information (Foissac *et al*, Current Bioinformatics 2009) was developed.
- A functional annotation pipeline, based on (i) various methods of patterns matching and motifs recognition, (ii) intracellular targeting prediction methods, and (iii) comparative genomics with other fungal genomes, was also developed.
- As major tool development, URGI proposes a pipeline for analysing and annotating repeats called REPET, known internationally for its high level of automation and accuracy. It was used within the framework of many international genome annotation projects. In particular, it produced transposable elements reference annotations for *Arabidopsis thaliana* and *Drosophila melanogaster* genomes (Quesneville & al PloS 2005, available at TAIR and Flybase the respective community genome repositories).

This platform was also used to annotate the *Botrytis cinerea* T4 genome sequence (40 Mb, 118 scaffolds). Other fungal genome annotation projects are under progress: *Leptosphaeria maculans* (Collaboration URGI/BIOGER and university of Melbourne) and *Blumeria graminis* (collaboration URGI-BIOGER and Imperial College, London).

Our annotation system allows the distributed annotation of genome sequence. Apollo is the graphical annotation editor allowing curators to change the gene structures according to various evidences (transcript, short-reads, protein similarity, and comparative genomics). Manual annotations (*i.e.* gene curation) are saved in a dedicated Chado database and shared with other members of the annotation community. When validated, genes/pseudogenes curated models are committed in a second database publicly accessible by Gbrowse.

A reference data warehouse to promote data exchange:

The great strength of the platform is also its capacity to centralize in the same site, different types of data and to link them in a reliable database system (with backup robots, a secure fire proof server room, and qualified staff for its maintaining). For many national and international projects, the platform is the reference data warehouse and is used to disseminate projects data to the whole scientific community.

The platform was chosen by the international grapevine consortium (IGGP) to manage grapevine genomic

annotations and to help the community through support, tools and databases development, to perform the manual gene annotation. It also hosted wheat genomic and genetic data for the European wheat scientific community and is used for the annotation of the first wheat chromosome sequence.

Conclusions

Data integration of sequences from the next generation sequencing technologies is a new scientific challenge in bioinformatics. Information systems such as GnpIS will play a very central role in the analysis of the ever increasing flow of new genomic data. Indeed, the need for data integration is increasing. Huge amount of data are now produced at lower costs than with previous technologies. This allows each lab to produce large amounts of data rapidly. Plant or fungal genome sequencing will be soon possible at a laboratory level. Moreover, more high-throughput experiments can be performed in every labs. This leads to a need of dedicated bioinformatics analysis tools to assemble reads, detect SNPs and map reads on a reference genome, but also tools to store, query and mine these data. Storage is not the only difficulty. Fast data access through queries is more challenging as this is critical for researchers to efficiently work with their data. To face this new challenge, we have developed the GnpIS architecture.