# Using semantic web technology to accelerate plant breeding.

Pierre-Yves Chibon[1,2,3], Benoît Carrères[1], Heleena de Weerd[1], Richard G. F. Visser[1,2,3], and Richard Finkers[1,3]

[1] Wageningen UR Plant Breeding, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands.
[2] Experimental Plant Sciences, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands
[3] Centre for BioSystems Genomics, Wageningen, 6708 PB, Wageningen, The Netherlands

**Abstract.** One goal within plant breeding is to find the causal gene(s) explaining a given phenotype. Semantic web technology brings opportunities to integration data and information accross spread data sources. Chebi2gene and Marker2sequence are two applications relying on this semantic web technology to integration genes, proteins, metabolites, pathways, literature. Their web-based interface allows biologists to use and explore this network of information.

**Keywords:** Semantic web applications, Data integration, Plant Breeding

## 1 Introduction

"Producing 70 percent more food for an additional 2.3 billion people by 2050 while at the same time combating poverty and hunger, using scarce natural resources more efficiently and adapting to climate change are the main challenges world agriculture will face in the coming decades" (http://www.fao.org/news/story/en/item/35571/). Plant breeding is part of the answer to this challenge. The FAO itself recognize it: "Plant breeding techniques can lead to improved crop varieties that increase yields, decrease losses" (http://www.fao.org/news/story/en/item/35686/). In order to improve crop varieties, breeders introgress genes of interest from one accession to another. The challenge is to pinpoint the gene(s) responsible for the improved traits. To find these genes, plant breeders use all the new types of information, which become available using high-throughput technologies, such as next-generation sequencing technology, RNASeq, proteomics and/or metabolomics.

As a daily practice, plant breeders associate these large datasets to one or several regions of the genome using advanced statistical methodology. These regions are called Quantitative Trait Loci (QTLs) and can be introgressed from one variety to another with the goal to developed an improved variety. However, a typical QTL region may contain over hundreds of genes, including genes negatively influencing the breeding goals. Complete genome sequences of many crop

plants are becoming available, including the genome of important food crops such as tomato [1] and potato [2]. The availability of structural and functional genome annotations makes it possible to investigate the QTL region for genes positively or negatively influencing the trait of interest.

Plant breeding, as most research areas nowadays, faces the problem of spreading data resources. Most plant species have their own website, ideally cross-referenced to major cross-species database such as UniProt or GO, but the number of resources available keeps increasing every year. When a researcher starts to investigate the genes in a specific QTL region on a genome, he will have to browse through an increasing number of websites and databases to collect and integrate information about each of these genes. One solution to this problem is to use semantic web technologies to aggregate and integrate the data from different resources in a way that would be and automated and expandable to new resources as they become available.

Within Wageningen UR Plant Breeding, we have developed two new tools relying on semantic web technologies to help breeders face this challenge, namely: Chebi2gene and Marker2sequence.

## 2 Materials and Methods

Our tools have been primary developed for usage with the Tomato genome sequence, however, they should are implemented in a universal manner. The limiting factor is that the annotation of a genome sequence should be available in RDF. To convert the Tomato genome annotation into RDF we developed a simple tool: gff2RDF, which retrieves and parses the gff file and outputs a RDF document of the annotation. This tool is used for Tomato and is being extended to work with Potato and Arabidopsis thaliana. In the conversion the gene annotation linked to external database have been converted to use the URI of these database, i.e.: the GO terms associated with the genes are identified using the URI provided by the OWL file of the Gene Ontology consortium, and similarly for the protein identifier against UniProt. The resulting RDF file has been uploaded into a Virtuoso OSE server (v 6.1.3) together with the RDF files provided by EBI for UniProt-core, UniProt-pathways, UniProt-go and UniProt-citations (all version 2011_10), CHEBI (version 2011_09), Rhea (release 33), the Gene Ontology OWL file as provided by the Gene Ontology Consortium (version 2011_11_03). Each resource is stored in its own graphs (Fig 1), allowing easier upgrades, and SPARQL is used to do the integration.

## 3 Chebi2gene, linking metabolites to genes

Metabolites are chemical compounds produced by an organism, in plants their actions can be related to traits of major interest such as flavor or disease resistance. The link between a metabolite and the genes involved in its expression is not always straightforward. Chebi2gene is a proof of concept allowing breeders to go from one metabolite to the associated gene(s). This allows biologists to

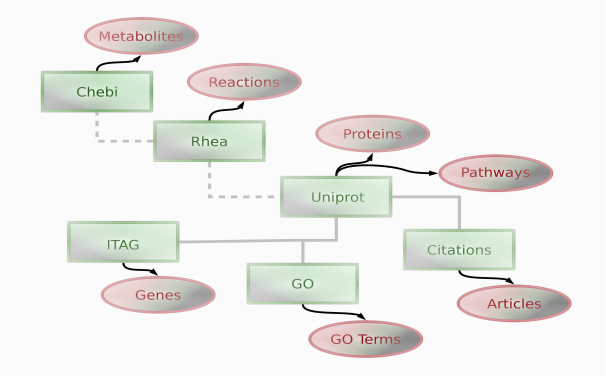


**Fig. 1.** This figure represents the integration of the different resources in our triple store. The green box defines the different RDF graphs and the red ellipse the type of information we extract from these graphs. The dashed gray line between around Rhea is for the fact that Rhea do not reuse the URI of Chebi or UniProt when refering to Chebi compound or a UniProt protein. The mapping is then indirect, as opposed to the plain gray line where the URI are consistent and shared.

find all the genes in the genome related to a metabolite. To find these associations, it uses CHEBI, Rhea, and UniProt databases and our RDF version of the tomato genome annotation. The input is either a CHEBI identifier or the name of the metabolite. If the input is a name, Chebi2gene will search the CHEBI database for all compounds having this name in their name and optionally in their synonyms. Once the metabolite has been uniquely identified with a CHEBI identifier, Chebi2gene search in Rhea all the chemical reactions in which it is involved, then all the proteins which are involved in these reactions and finally all the genes from the genome annotation which are related to these proteins. These searches are performed using SPARQL queries on our Virtuoso server across the different graphs of the different resources.

Chebi2gene is available at: http://www.plantbreeding.wur.nl/chebi2gene

For example, when searching for "beta-carotene" in chebi2gene, three molecules containing "beta-carotene" are returned: "beta-carotene", "beta-carotene 5,6-epoxide", and "(5S,6R)-beta-carotene 5,6-epoxide". From these three molecules, the first one is the molecule of interest. It has the CHEBI identifier 17579. Searching with this identifier in chebi2gene, we can find that this compound is involved in 4 reactions which are associated with 10 proteins. Amongs these proteins is "Lycopene beta cyclase" (UniProt ID: Q38933) which is associated with two pathways: "Carotenoid biosynthesis; beta-carotene biosynthesis" and "Carotenoid biosynthesis; beta-zeacarotene biosynthesis" and four genes: Solyc04g040190.1.1 (chromosome 4) and Solyc10g079480.1.1 (chromosome 10) which are both "Beta-lycopene cyclase" and Solyc06g074240.1.1 (chromosome

6) and Solyc12g008980.1.1 (chromosome 12) which are both "Lycopene beta cyclase".

## 4 Marker2sequence explore a genome region for a candidate gene

Marker2sequence [3] aims at mining quantitative trait loci (QTLs) for candidate genes. For each gene, within the QTL region, marker2sequence uses semantic web integration technology to integrate putative gene function with associated Gene Ontology terms, proteins, pathways and literature. This integration is performed using SPARQL queries against our triple-store. As mention earlier, a typical QTL region easily contains several hundreds of genes, this gene list can then be further filtered using a keyword based query on the aggregated annotations. This single query search for the given keyword in the gene annotation and GO terms, proteins and literature associated with this gene. More precisely, it searches the keyword in he name and definition and synonym of the Gene Ontology term associated with the gene. It searches the keyword in the protein name and description of each protein associated with the gene. It searches the keyword in the pathway name of each pathways associated to these proteins and finally it searches in all title and abstract of the literature associated with these proteins. If any of these elements contains the searched keyword the gene is selected as a potentially interesting gene and returned to the user Marker2sequence will help breeders to identify potential candidate genes for their traits of interest.

Marker2sequence is available at: http://www.plantbreeding.wur.nl/BreeDB/marker2seq

For example, $\beta$-carotene content is a trait influencing the color of tomatoes [4]. Based on our QTL analysis, using data from the Solanum lycopersicum x Solanum galapagense LA0483 RIL population [5], this compound has one QTL on chromosome 6 (between TG253 and TG314). Marker2sequence identifies 988 genes in this region of the chromosome 6. A query with the keyword: beta-carotene, returns the gene Solyc06g074240.1.1. This gene, Solyc06g074240.1.1, is associated with the GO term for carotenoid biosynthetic process, the pathway for Carotenoid biosynthesis and more specifically the part of beta-carotene biosynthesis. Information for each gene can be quickly mined using Marker2sequence and this gene is the candidate for our trait of interest.

## 5 Conclusion

Both Chebi2gene and Marker2sequences are tools presenting the potential of semantic web integration for the plant breeding domain.

All the tools presented in this abstract have been licensed under free license and are available at: http://github.com/PBR/

## References


[1] The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485** (2012) 635641.

[2] The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. Nature **475** (2011) 189195.

[3] Chibon, P-Y., Schoof, H., Visser, R.G.F., Finkers, R.: Marker2sequence, mine your QTL regions for candidate genes. Bioinformatics **28** (2012) 1921-1922.

[4] Lincoln, R.E., Porter, J.W.: Inheritance of beta-carotene in tomatoes. Genetics **35** (1949) 206-211.

[5] Paran, I. Goldman, I., Tanksley, S.D., Zamir, D.: Recombinant inbred lines for genetic mapping in tomato. Theor. Appl. Genet. **90** (1995) 542-548.