GNPannot

A community annotation system applied to sugarcane sequences

cirad

Olivier GARSMEUR
ICSB workshop

South Green bioinformatics platform

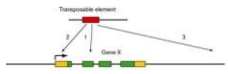
International Plant & Animal Genome XXI January 12-16, 2013 - San Diego, CA, USA

Introduction to concepts and methods

Two main concepts:

Identify the different elements of the genome, (location and stucture) :

1 - structural annotation



Attribute a biological information to these elements : **2** - functional annotation

Automatic annotation – Gene prediction

Intrinsec methods (ab-initio)

- Only based on computational analysis using statistical models.
- Probabilistic models like *Hidden Markov models* (HMM) for discriminating coding and non-coding region of the genome.

The sequences of exons, splice sites and introns have different statistical properties:

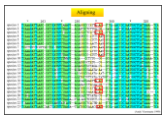
- GC% → Introns are AT rich and
- splicing site is almost GT / AG (for plants 95%)

-Need a training set genes. (learning)

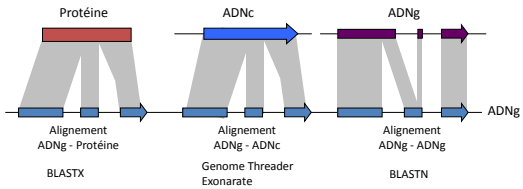
The learning set is composed of several hundred of gene-sequences manually annotated derived from cDNAs / genomic alignments. Ideally, these genes represent the diversity of the genes that can be found in the genome.

Automatic annotation – Gene prediction

Extrinsec methods



Based on comparative approaches
= sequence similarities
The sequence to annotate is compared with databases.



Automatic annotation : Extrinsic methods

conserved protein domains = signatures

database of predictive protein "signatures" can be used for the classification and automatic annotation of proteins.

Interproscan classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains and important sites.

Domain databases used by interproscan:

Prosite patterns	GENE3D
Pfam	HAMAP
ProDom	PANTHER
Superfamily	PIRSF
TIGRFAMs	

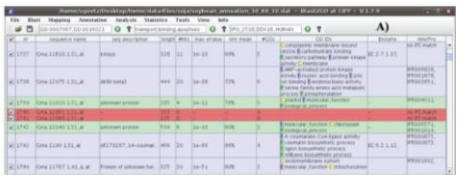
Automatic annotation : Pipeline for functional annotation

blast2go

http://www.blast2go.com/b2ghome

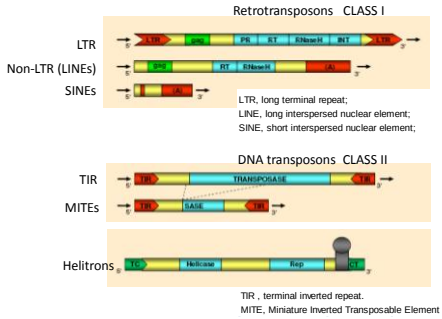
Blast2GO is a bioinformatics tool for functional annotation of sequences, primarily based on the gene ontology (GO) vocabulary.

Transfers the function from homologous sequences through an elaborate algorithm that considers similarity, the extension of the homology, the database of choice, the GO hierarchy. Its also supports InterPro, enzyme codes and KEGG pathways.



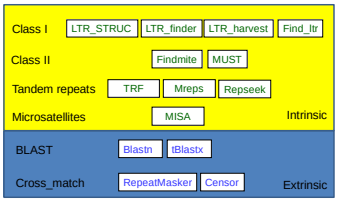
Transposable Elements (TEs)

Genes represent only a little part of the genome. Some regions can be gene-rich but some other can contain a majority of repeated elements.



Annotation of transposable elements : tools

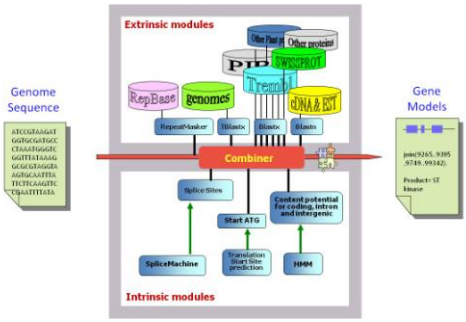
-Several ab-initio programs can be used to detect structures of TEs = intrinsec
-Comparisons with databases can be used to classify the elements = extrinsec



- Pipeline to detect and classify Tes = REPET pipeline (Flutre et al, 2011 PLOS one)

Integrative method – The combiner

Integrative methods = combine ab-initio (intrinsec) + comparative approaches (extrinsec) improving significantly the annotation.



GNPAnnot Project

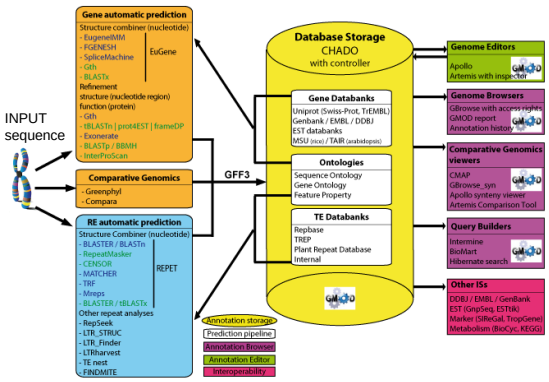
GNPAnnot is a community system for structural and functional annotation dedicated to plants, insects and fungus genomes allowing both automatic predictions and manual curations of genomic objects.

The system is currently being used for various plants, insect and fungus species.

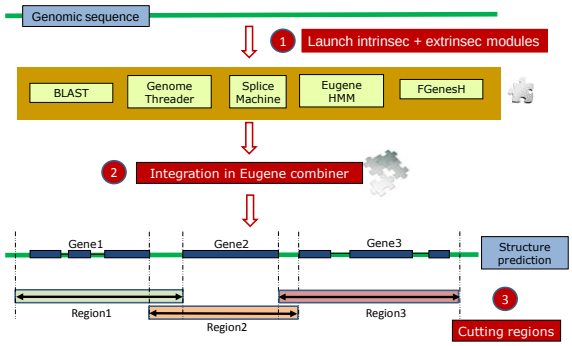
Each specie has a personalized pipeline

http://southgreen.cirad.fr/

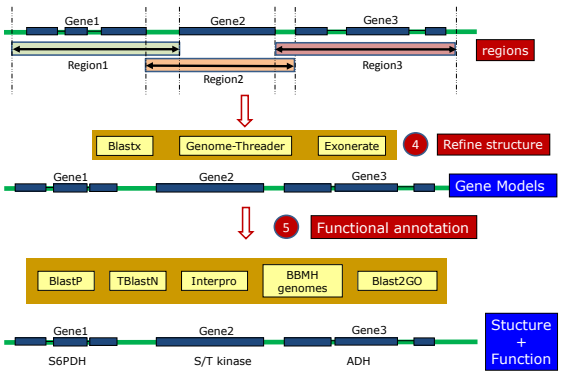
GNPAnnot platform



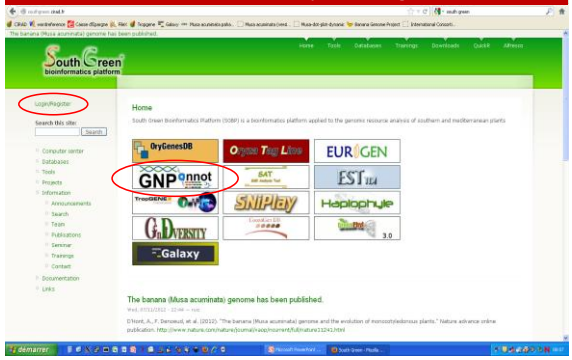
Sugarcane Pipeline



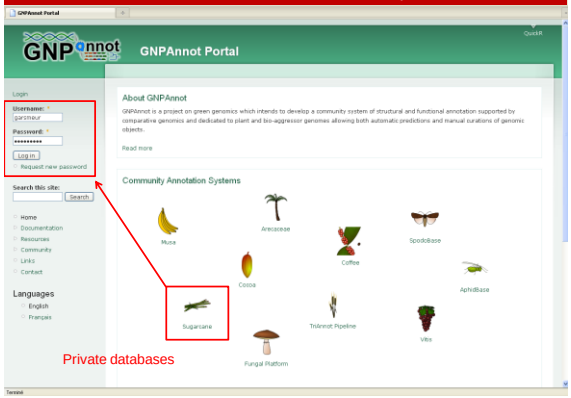
Sugarcane Pipeline



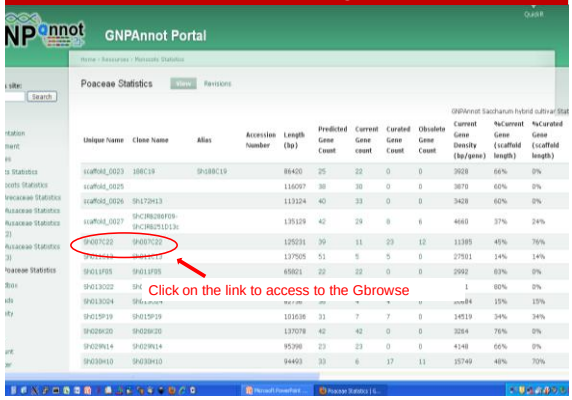
GNPAnnot on the web : http://southgreen.cirad.fr



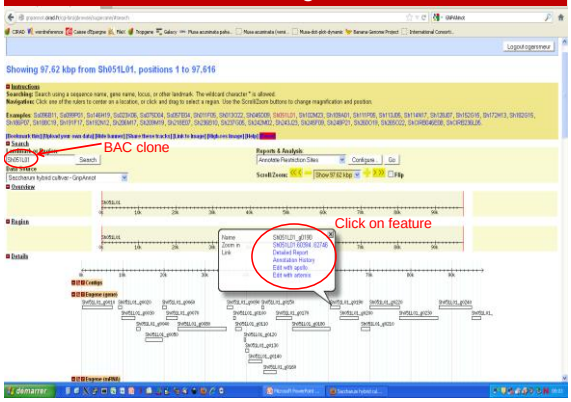
GNPAnnot Portal, secure access for private data



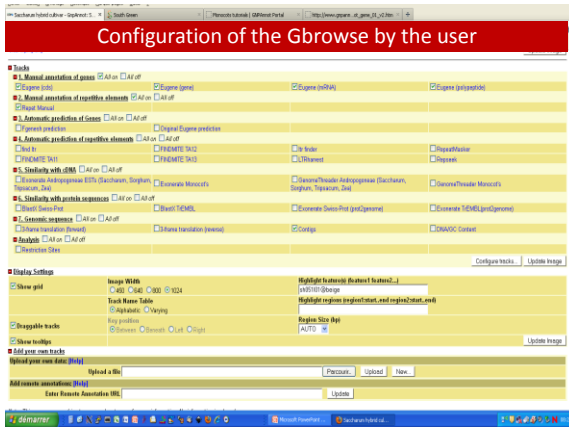
Tables with links to gbrowse



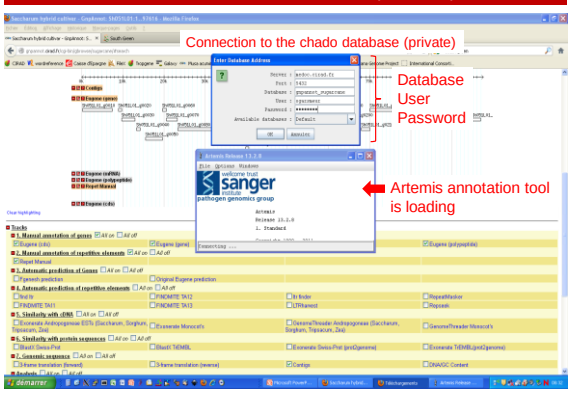
interactive gbrowse



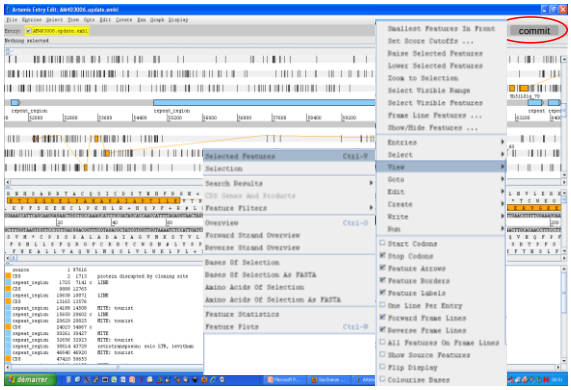
Configuration of the Gbrowse by the user



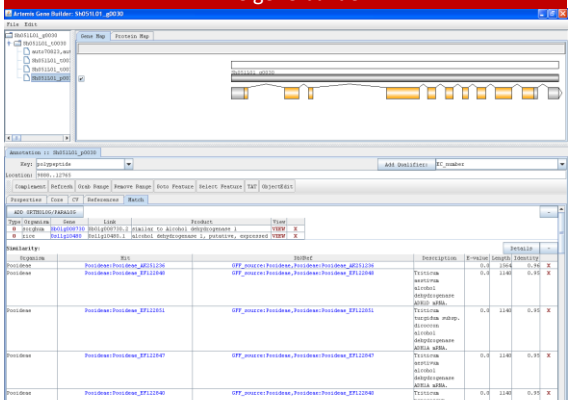
From the Gbrowse , click and edit your sequences



artemis : to see, to check, to correct, to validate, to store



The gene builder



History report

A web site for sugarcane genomic resources

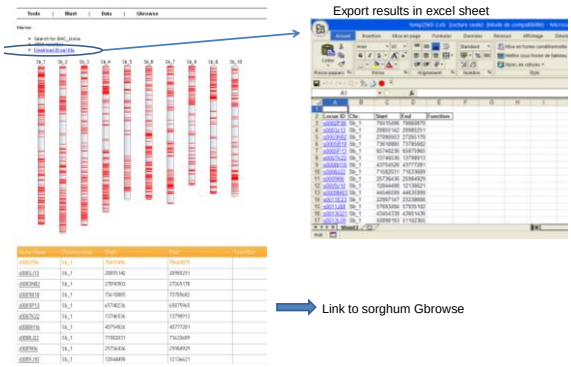
Sugarcane sequences (BACs and other sequence data, markers, DArTs ...) could be available from a private web portal.
Sugarcane sequences will be annotated with GNPAnnot and anchored on the sorghum sequence

example

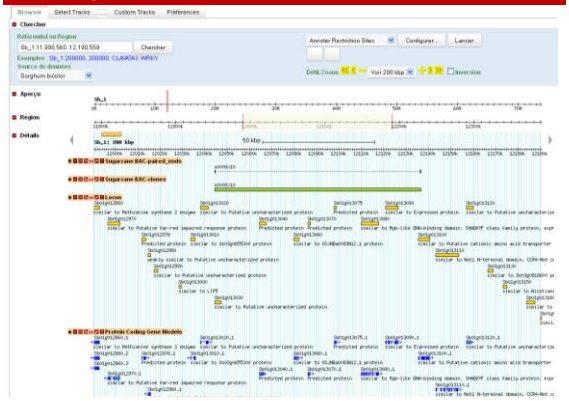
R570 BAC-ends from A. Paterson anchored on sorghum

Sugarcane BAC clones sequences : links to GNPAnnot to see annotation , and to Sorghum GBrowse

Location of the R570 BAC-end sequences on the sorghum



Sorghum Gbrowse with R570 BAC clone locations



Link to sugarcane GNPAnnot Gbrowse with BAC annotations