Annotation of genomic sequences
An example on 1 region of the rice chromosome 1

1) TD objectives

The objective is to identify the TD, a large genomic region, all the coding structures (genes) using a set of inherent annotation methods *(ab initio* prediction) and extrinsic (making use existing databases). An automatic annotation pipeline available on the platform galaxy will be used to perform automatic annotation of a region of rice chromosome 1. Comparison of automatic annotations obtained with different tools leaves sometimes appear to differences in the number of predicted genes and in their structure. The use of Artemis will view the automatic annotation results and highlight the differences. You will use Artemis to perform manual correction of the structural annotation. Beyond structural information of the genomic region under consideration, it is possible to acquire functional information by interviewing databases and searching for sequence similarity and conserved protein domains (Signatures). The function of the predicted genes will be allocated more or less confidence in function of the Significance of results alignments with known genes or proteins.
The pipeline (workflow galaxy) that we will use for the annotation includes the following modules:

Intrinsic methods

Splicemachine http://bioinformatics.psb.ugent.be/webtools/splicemachine/ predicted splice sites by the use of the learning method called "linear support vector machine" (LSVM, http://fr.wikipedia.org/wiki/Machine_à_vecteurs_de_support) using models derived from the genome from *Arabidopsis thaliana* or the human genome.
EugeneIMM uses the method IMM **(Interpolated** Markov Modeler) for discriminating the coding regions of the noncoding.
FGENESH http://www.softberry.com/berry.phtml is a gene prediction method based on methods statistical HMM (hidden Markov chains) with a supervised learning phase.

Extrinsic methods

BLAST (Basic Local Alignment Search Tool) http://www.ncbi.nlm.nih.gov/BLAST/. The program compares nucleotide or protein sequences (depending on the type of BLAST used) and calculates the significance results (based on the percent identity and length of the game). Translated **BLASTX** sequence in the 6 phases and compares it to a database "protein" kind Swissprot or Trembl.
BLASTP compares the sequence "protein" in a database "protein" kind Swissprot or Trembl.
TBLASTN compares the "protein" in sequence databases "nucleotide" translated in the 6 phases, NR type (non-redundant sequences), EST (Expressed Sequence Tag) or complete genomes.
Genome Threader http://www.genomethreader.org/ predicted gene structures through similarities cDNAs or ESTs and / or aligned protein sequences (consensus alignments, taking into account splicing). It uses an intron excisor and a model "Baysian Splice Site Models" (BSSMs) to identify the exon-intron boundaries.
Exonerate http://www.ebi.ac.uk/guy/exonerate/ is an alignment tool sequences pairs. There is able to consider different models of alignment including the ability to align an EST against a genomic sequence or a protein sequence against a genome.

Combiner

EuGène (http://eugene.toulouse.inra.fr/) is a tool for integration of previous modules in the process annotation. It outputs a maximum score of prediction, that is to say the more consistent as possible with the information provided by each of the modules.

2) Automatic annotation pipeline execution (Workflow Galaxy)

The pipeline used to perform automatic annotation has already been initiated (to save time, as it takes about 1 hour calculation time for the entire pipeline). It consists of the modules described above which will be executed one after the other, or in parallel. The programs are interconnected and is in general the automatic annotation pipeline.

Description of the workflow:

For structural annotation (Figure 1), three bricks are used: "SpliceMachine" and "Eugene" (including EuGeneIMM). The result of an analysis performed with FGENESH is also filled with "Eugene" after format conversion ("GNPAnnot Converters: FGENESH")

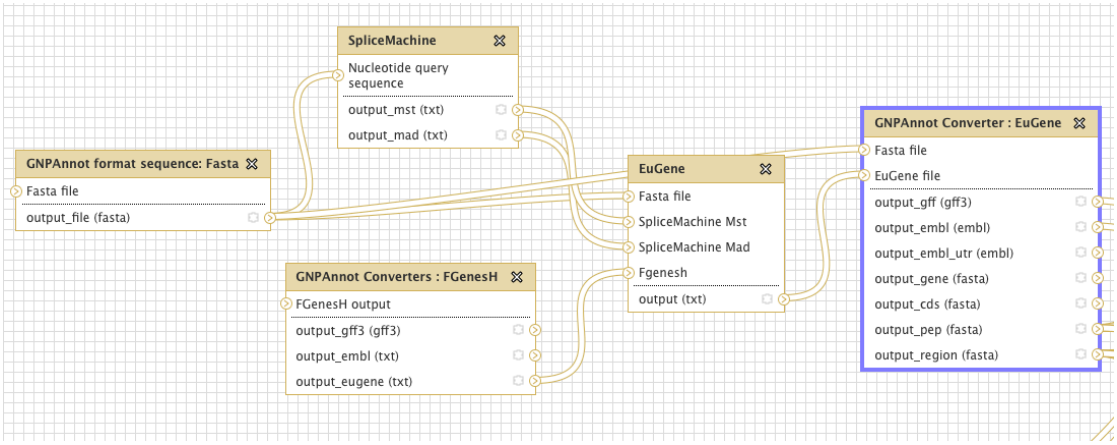

Figure 1: Workflow Galaxy for structural annotation of genomic sequence

The resulting file, "Eugene output 'is gross output" Eugene ". It serves as a starting point functional annotation and other steps to refine the predicted gross structure. Brick "GNPAnnot Converter Eugene" makes it possible to extract files containing the structure of the predicted genes (Gff3 and EMBL) and the fasta files necessary for functional annotation. Brick "Eugene" produces the following output files:

- Eugene output without functional annotation (gff3 and emblem)
- Sequences of the predicted genes including introns (fasta)
- Sequences of CDS (Gene Coding Sequence without introns) (fasta)
- Protein sequences (fasta)
- Extracts regions around the predicted genes, to refine the structure (fasta)

Functional annotation

To assign a function to a gene predicted by Eugene (Figure 2), brick "GNPAnnot Converter Blastp" combines the results of several sources of BLAST (SwissProt MSU Rice genome annotation project = Rice MSUv6.1, Proteome Sorghum extract the database Phytozome) and transfers the function of the protein more similar so identified

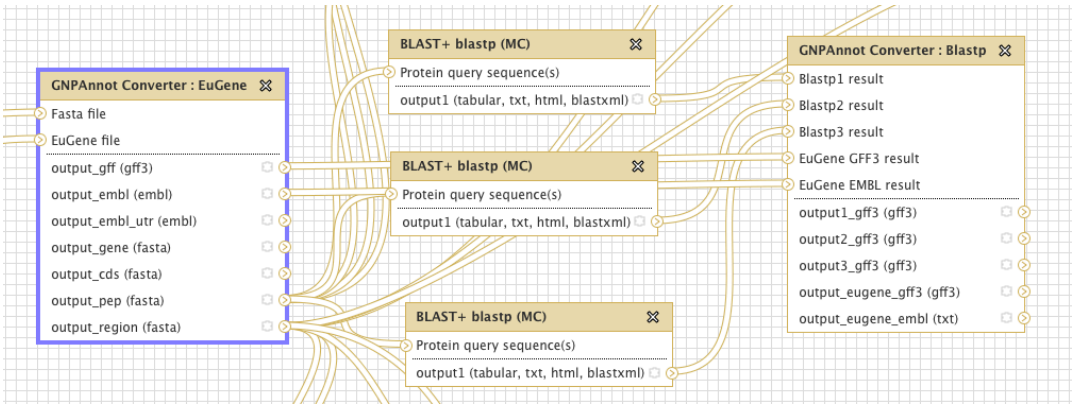

Figure 2 : Workflow Galaxy for functional annotation

Development of structural annotation:

To clarify the structure of the predicted genes (Figure 3) is used in a first time a combination of TBLASTN and exonerate the rice EST databases *(Oryza sativa* and *Oryza glaberrima)* and sorghum. Also used in parallel combination BLASTX / exonerate and Genome Threader program on the extended nucleic sequence between genes (Figure 4).

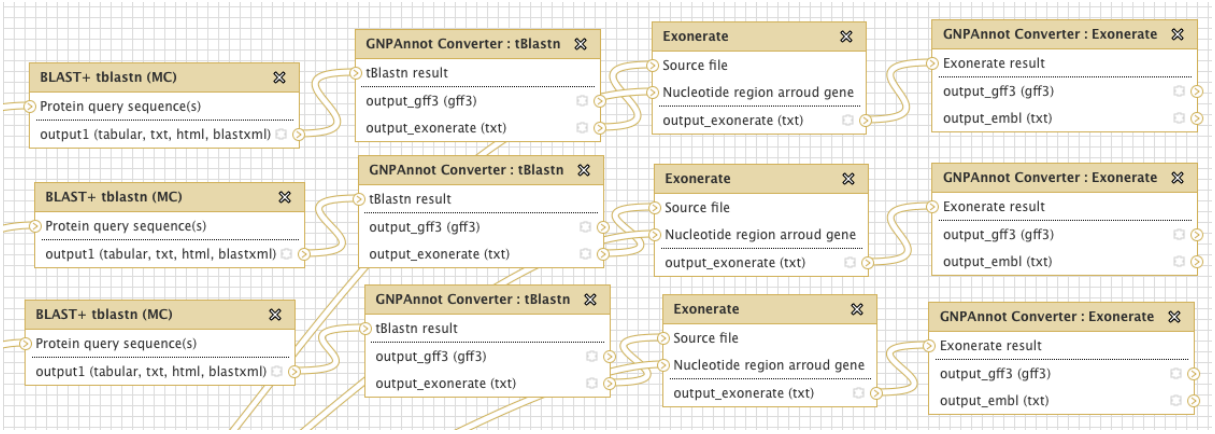

Figure 3 : Workflow Galaxy to improve functional annotation

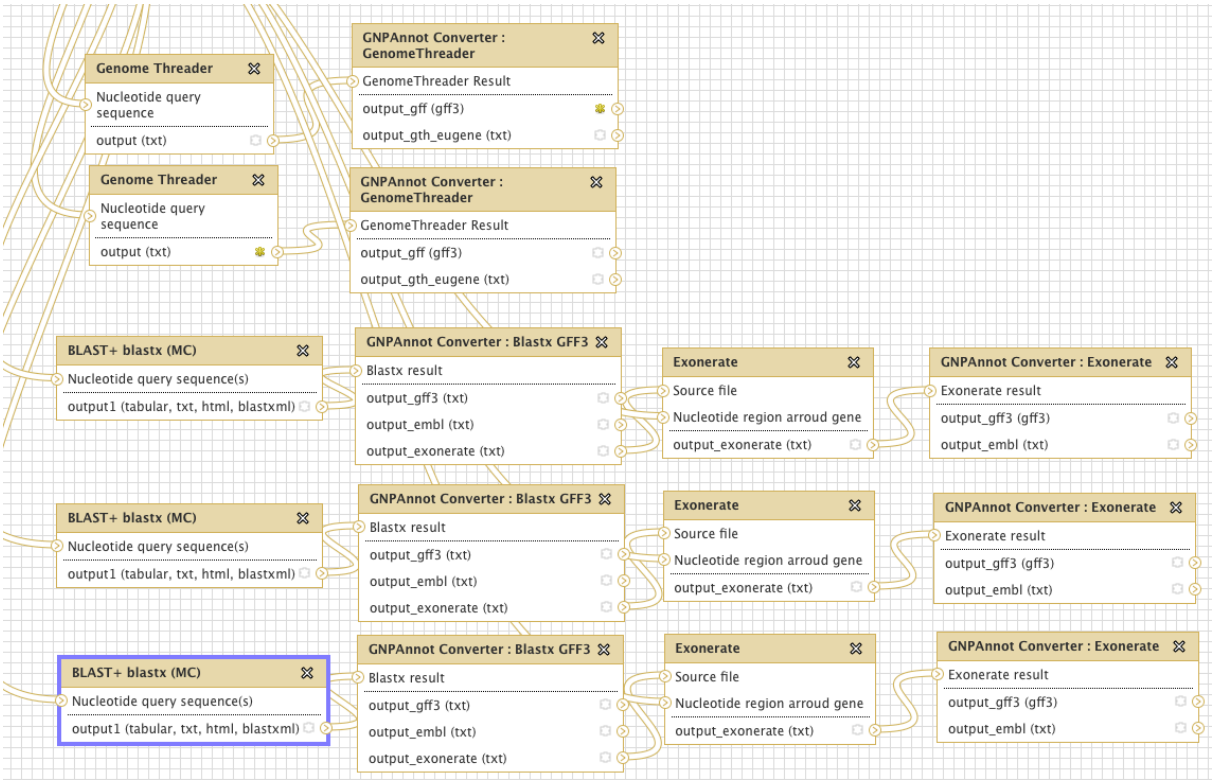

Figure 4: Workflow Galaxy to improve the structural annotation from extended nucleic sequences genes.

---

* Recovery workflow output files:
Take the following output files in the history galaxy:

•
(8) FGENESH (emblem) at EMBL Software File size FGENESH

•
(44) Eugene (EMBL): File format EMBL Eugene program

•
(64) exonerate OG_ngs (EMBL): EMBL file corresponding to the combination of programs TBLASTN / exonerate the contigs Rice (ssp. glaberrima)

•
(66) exonerate OS_mrnas (EMBL): EMBL file corresponding to the combination of programs TBLASTN / exonerate the bank IS Rice (ssp japonica)

•
(62) exonerate SB_mrnas (EMBL): EMBL file corresponding to the combination of programs TBLASTN / exonerate the bank IS sorghum.

•
(72) Rice exonerate (EMBL): EMBL file corresponding to the combination of programs Blastx / exonerate the proteome of Rice (MSU 6.1)
•
(70) exonerate SwissProt (EMBL): EMBL file corresponding to the combination of programs Blastx / exonerate the bank UniProtKB / SwissProt
•
(68) exonerate Sorghum (EMBL): EMBL file corresponding to the combination of programs Blastx / exonerate the proteome Sorghum
•
(3) Os01_36429_36558.fna.repeat: File format corresponding tabbed BLAST to compare the sequences with repeated sequences databases

3) automatic annotation visualization with "Artemis" and manual correction of the annotation

à◊ Go to the website of "Artemis" (sanger).
http://www.sanger.ac.uk/resources/software/artemis/
à◊ Click the "download" tab.
à◊ Click the "launch" button to start the download of the application "Artemis"
Once "Artemis" downloaded, the application can be run by double-clicking the downloaded file.
The execution of Artemis requires prior installation of Java on your computer (already done for TD today).

Open the automatic annotation file "Eugene" in format EMBL artemis:

à◊ Click the File / Read An Entry menu
Files of type: All files
File name: File name:. Galaxy ___- [EuGene_ (emblem)] txt
To the question "There Were warnings while reading - view now? "No answer (if you click yes, this is okay, but you'll warnings (warnings) on the format of annotations)

Q1: What is the identifier of the first predicted gene (locus tag =)?
Q2: How are they coding structures predicted by Eugene?

Creating a track (new entry) that correspond to the manual annotation layer:

à◊ Click the Create / New Entry menu
By default, your new track called "no name". To rename your new track:
à◊ Click the Entries / Set Name Of Entry / no name menu
A dialog box opens. Type the name you want to give to the track. For example "Manual_annotation".
Now copy the features of the Galaxy ___- track [EuGene_ (emblem)]. Txt in your new track. To be sure to copy only the features of this entry,
à◊ Check the Entries menu that only the Galaxy ___- track [EuGene_ (emblem)]. txt is checked. In the case otherwise uncheck the other tracks.
à◊ In the window that lists the features (the one containing the gene information, CDS as text)

select all features (click in the window and [Ctrl] + A)
à◊ Click the Edit menu / Copy Selected Features To / Manual_annotation
This annotation track will be used to carry out the changes. It is the annotation track manual.

Open other files in "Artemis" to add additional information layers:

File name: File name:. Galaxy ___- [FGenesH_ (emblem)] txt
File name:. Galaxy ___- [Exonerate_OG_ngs_ (EMBL)] txt
File name:. Galaxy ___- [Exonerate_OS_mrnas_ (EMBL)] txt
File name:. Galaxy ___- [Exonerate_SB_mrnas_ (EMBL)] txt
File name:. Galaxy ___- [Exonerate_Rice_ (EMBL)] txt
File name:. Galaxy ___- [Exonerate_Sorgho_ (EMBL)] txt
File name:. Galaxy ___- [Exonerate_SwissProt_ (EMBL)] txt
File name: Galaxy ___- [Os01_36429_36558.fna.repeat]
Note: If you want to remove an entry: Menu Entries / Remove An Entry / select the file to remove
* To facilitate the visualization of the results:
à◊ Right click in the viewport annotations
Check à◊ One Line Per Entry
Uncheck à◊ Feature Labels

Comparison of the predicted structure by Eugene and that predicted by FGENESH

à◊ Menu Entries / check the box next to:. Galaxy ___- [FGenesH_ (emblem)] txt
Q3: What is the major difference between the prediction and the EuGène FGENESH?
Q4: What can this be due?
Help yourself to the information contained in the various files you have open in "Artemis" in checking the box next to the track you want to display (Menu Entries).

Use the comparison data with the bank type databases EST / cDNA

à◊ Menu Entries / check the box next to:
Galaxy ___- [Exonerate_OS_mrnas_ (EMBL)]. Txt (TBLASTN / exonerate the bank IS Rice (ssp japonica)
Galaxy ___- [Exonerate_SB_mrnas_ (EMBL)]. Txt (TBLASTN / exonerate the bank IS sorghum)
NB: Artemis is now able to read the mapping files (file type bam) corresponding to RNAseq data type of alignment with the genomic sequence
Q5: What are the main differences between the structure prediction EuGène and those reconstructed by "Exonerate"?
Q6: In light of this information, what structure you seem most consistent FGENESH or Eugene?

Use the annotation data of species or subspecies close

à◊ Menu Entries / check the box next to:
Galaxy ___- [Exonerate_Sorghum_ (EMBL)]. Txt (BLASTX / exonerate against proteome sorghum)
Galaxy ___- [Exonerate_OG_ngs_ (EMBL)]. Txt TBLASTN / exonerate the contigs Rice (ssp. Glaberrima)
Q7: From the set of comparisons, which are the structure points need to check? What types data can help you make a decision?

Question annex the region of rice chromosome analyzed:
Q8: Is some information allows you to establish a micro-synteny relationship with sorghum?

Use protein databases (Swissprot)

à◊ Menu Entries / check the box next to:
File name: Galaxy___ Galaxy ___- [Exonerate_SwissProt_ (EMBL)] txt (BLASTX / exonerate against base. protein data UniProtKB / Swissprot
Q9: Are the results allow you to confirm certain structural elements to correct?

Correct a gene structure

Changes will be made on the features of the manual annotation track you have created.
To add / edit an exon in a CDS
à◊ click the CDS correct then [Ctrl] + E
An edit window opens. Exons coordinate information is listed in the box "Location":
join(1124..1306,1910..1981,2081..2184,2265..2367,2455..2616,2741..3018,6126..6276,6493..6657,6775..6852, 7158..7244,7334..7420,7477..7588,7731..7834,8133..8208,8652..8802,8944..9012,9106..9172,9339..9383
Add à◊ the coordinates of the new exon and confirm by clicking the "Apply" then "OK"
Check à◊ junction GT / AG of exons created
Q10: In your opinion what are the correct details of the first exon of the gene?
à◊ Check with the information you have the structure of the first exon.
à◊ Change and check the connections at the splice sites.
Q11: In your opinion what are the correct details of the last exon of the gene?
à◊ Check with the information you have the structure of the last exon.
à◊ Change and check the connections at the splice sites.

Validation corrections structural annotation

To check whether structural changes have improved (or not) the annotation of the gene, you will compare the initial annotation of Eugene (Galaxy ___- [EuGene_ (emblem)]. txt) to your manual annotation.
For that we will align the predicted proteins by each of the two annotations with basic proteins of Uniprot / Trembl data.
Recover the protein sequence of the first gene in Eugene track and on the track manually annotated
à◊Clic right on CDS object
à◊View / Amino acids of selection as fasta
Copy / paste the sequence in a text editor (notepad) and save each sequence.
For example: locus_tag_ori.faa (for sequence Eugene) and locus_tag_cor.faa (for your annotation corrected).
à◊ Launch a browser, open two tabs and go to the following address
http://www.expasy.ch/tools/blast/
or http://www.uniprot.org/ Blast tab
à◊ Copy and paste the sequence locus_tag_ori.faa in one of the tabs and the locus_tag_cor.faa the other.
Launch BLASTp by clicking the Run button BLAST

Q12: Observe the alignments, is your best annotation that the initial automatic annotation?
What clues help you conclude?
What additional changes in gene structure would you bring?
à◊ Refine your gene structure and check the terminals to the splice sites.

Functional annotation of the gene
Run "InterProScan" for research conserved protein domains and try to identify the family and
Depending on your protein
Launch a web browser and go to the website: http://www.ebi.ac.uk/Tools/pfa/iprscan/
Recover the protein sequence of the first gene of your manual annotation track
à◊Clic right on CDS object
à◊View / Amino acids of selection as fasta
à◊Copier and paste the sequence in the window "InterProScan"
à◊Lancez search Interpro.
à◊ Start a browser and go to the following address: http://www.expasy.ch/tools/blast/ or
http://www.uniprot.org/ Blast tab
à◊ Copy and paste the protein sequence of the first gene of your manual annotation track.
Start the BLASTP by clicking the Run button BLAST
Analyze your alignments blastp against Uniprot and be the score of "InterProScan"
Q13: What is the Uniprot accession for your gene?
Q14: What are the conserved domains that "InterProScan" has detected.
Q15: With these elements set the putative function of your gene using a controlled vocabulary.
à◊Cliquez on the CDS and [Ctrl] + E to open the edit box of the gene to annotate.
à◊Vérifiez the putative function of the gene is knowledgeable in qualifying the "product"
Add qualifiers and inform:
/ Evidence curated =
/ Status = "finished"
Annotation repeated elements
In the TD, you spotted the presence of repeated elements in an intron of the gene.
To delimit the repeat region and try to define the type of item you will recover this sequence
intron and make a similarity search against a database of repeated elements rice
à◊ Note the coordinates of the intron (beginning and end)
à◊Créez a new feature: Menu Create / New Feature
à◊Par default key "Key" inquired "misc_feature". Change to "repeat_region"
à◊Entrez the coordinates of the intron in the "rent": 3019..6125 and confirm ("Apply" then "OK")
à◊Clic right on repeat_region object
à◊View / bases of selection as fasta
à◊Copier and paste the sequence in the "Censor" window to the next adreese: http://www.girinst.org/censor/
à◊Selectionnez based TEs rice in the list of choices
à◊ Start the search
à◊ Note the coordinates and type of dectéctés elements.
à◊ modify your annotated accordingly by editing your feature "repeat_region"
-
Contact Information
-
Information in a call / note = "TE identified my family"
-
/ Annotator_comment = "enter the information you want to keep,
for example, how you go about identifying the TE "
Backup your manual annotation
To save your manual annotation:
à◊Menu File / Save As An Entry / format EMBL / Ma_piste_manuelle
Name the file and sauvergardez.
à◊Fermez artemis.
Q16: From concepts you learned in this TD, can you see the annotation
Automatic and manual of the analyzed area ?