Annotation de séquences génomiques
Exemple d'une région du chromosome 1 de riz

1) Objectif du TD

L'objectif du TD est d'identifier, sur une grande région génomique, l'ensemble des structures codantes (gènes) en utilisant un ensemble de méthodes d'annotation intrinsèques (prédiction *ab initio*) et extrinsèques (faisant appel aux bases de données existantes). Un pipeline d'annotation automatique disponible sur la plateforme galaxy sera utilisé pour réaliser l'annotation automatique d'une région du chromosome 1 du riz.
La comparaison des annotations automatiques obtenues avec différents outils laisse apparaitre parfois des divergences sur le nombre de gènes prédits et au niveau de leur structure. L'utilisation d'Artémis permettra de visualiser les résultats d'annotation automatique et de mettre en évidence ces différences. Vous utiliserez Artemis afin de réaliser la correction manuelle de l'annotation structurale. Au-delà des informations structurales de la région génomique considérée, il est possible d'acquérir des informations fonctionnelles en interrogeant les bases de données et en recherchant des similarités de séquences et des domaines protéiques conservés (signatures). La fonction des gènes prédits sera attribuée avec plus ou moins de confiance en fonction de la significativité des résultats d'alignements avec des gènes ou des protéines connues.
Le pipeline (workflow galaxy) que nous allons utiliser pour l'annotation comprend les modules suivants:

Méthodes intrinsèques

Splicemachine http://bioinformatics.psb.ugent.be/webtools/splicemachine/ prédit les sites d'épissage par l'utilisation de la méthode d'apprentissage dite « linear support vector machine » (LSVM, http://fr.wikipedia.org/wiki/Machine_à_vecteurs_de_support) à partir de modèles issus du génome d'*Arabidopsis thaliana* ou du génome humain.

EugeneIMM utilise la méthode IMM **(**Interpolated Markov Modeler) pour discriminer les régions codantes des non codantes.

FGenesh http://www.softberry.com/berry.phtml est une méthode de prédiction de gènes basée sur des méthodes statistiques HMM (chaines de Markov cachées) avec une phase d'apprentissage supervisée.

Méthodes extrinsèques

BLAST (Basic Local Alignment Search Tool) http://www.ncbi.nlm.nih.gov/BLAST/ . Le programme compare des séquences nucléotidiques ou protéiques (en fonction du type de BLAST utilisé) et calcule la significativité des résultats (basé sur le pourcentage d'identité et la longueur du match).

BLASTX Traduit la séquence dans les 6 phases et compare à une base de données « protéines » type Swissprot ou Trembl.

BLASTP compare la séquence « protéine » à une base de données « protéines » type Swissprot ou Trembl.

TBLASTN compare la séquence « proteine » à des bases de données « nucléotide » traduit dans les 6 phases, type NR (séquences non redondantes), EST (Expressed sequence Tag) ou des génomes complets.

Genome Threader http://www.genomethreader.org/ prédit des structures de gènes au travers de similarités avec des ADNc ou EST et/ou des séquences protéiques alignées (alignements consensus, tenant compte des épissages). Il utilise un exciseur d'introns et un modèle « Baysian Splice Site Models » (BSSMs) pour identifier les limites exons-introns.

Exonerate http://www.ebi.ac.uk/~guy/exonerate/ est un outil d'alignement de séquences deux à deux. Il est capable de prendre en compte différents modèles d'alignements avec notamment la possibilité d'aligner un EST contre une séquence génomique ou bien une séquence protéique contre un génome.

Combiner

EuGène (http://eugene.toulouse.inra.fr/) est un outil d'intégration des modules précédents dans le processus d'annotation. Il produit en sortie une prédiction de score maximal, c'est-à-dire la plus consistante possible avec les informations fournies par chacun des modules.

2) Execution du pipeline d'annotation automatique (Workflow sous Galaxy)

Le pipeline utilisé pour réaliser l'annotation automatique à déjà été lancé (pour gain de temps, car il faut environ 1 heure de temps de calcul pour l'ensemble du pipeline). Il se compose des modules décrit précédemment qui vont être exécutés les un à la suite des autres, ou en parallèle. Les programmes sont interconnectés et constitue dans leur ensemble le pipeline d'annotation automatique.

*Description du workflow :

Pour l'annotation structurale (Figure 1), 3 briques sont utilisées : « SpliceMachine » et « EuGene » (incluant EuGeneIMM). Le résultat d'une analyse réalisée avec FGenesh est également renseignée à « EuGene », après conversion de format (« GNPAnnot Converters : FGenesH »).

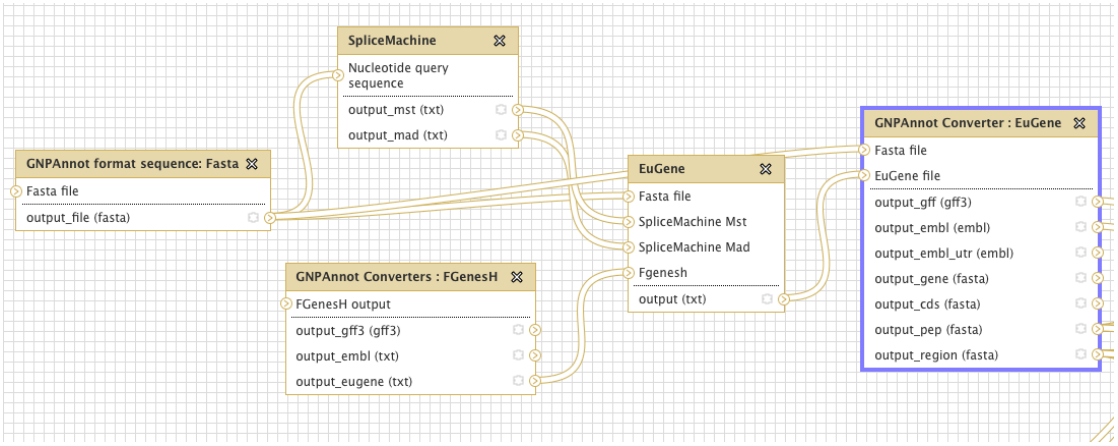

Figure 1 : Workflow Galaxy pour l'annotation structurale de séquence génomique

Le fichier résultant, « EuGene output », correspond à la sortie brute de « EuGene ». Il sert de point de départ à l'annotation fonctionnelle et aux autres étapes permettant de raffiner la structure brute prédite. La brique « GNPAnnot Converter : Eugene » permet en effet d'extraire des fichiers contenant la structure des gènes prédits (gff3 et embl) ainsi que les fichiers fasta nécessaires à l'annotation fonctionnelle.
La brique « EuGene » produit les fichiers de sortie suivants :

- EuGene output, sans annotation fonctionnelle (gff3 et embl)
- Sequences des genes prédits comprenant les introns (fasta)
- Sequences des CDS (Gene Coding Sequence, sans introns) (fasta)
- Séquences proteiques (fasta)
- Extrait les régions autour des gènes prédit, pour raffiner la structure (fasta)

Annotation Fonctionnelle

Pour attribuer une fonction à un gène prédit par EuGene (Figure 2), la brique « GNPAnnot Converter : Blastp » combine les résultats de plusieurs sources de BLAST (SwissProt, MSU Rice genome annotation project =Rice MSUv6.1, Protéome Sorgho extrait de la base de donnée Phytozome) et transfère la fonction de la protéine la plus similaire ainsi identifiée.

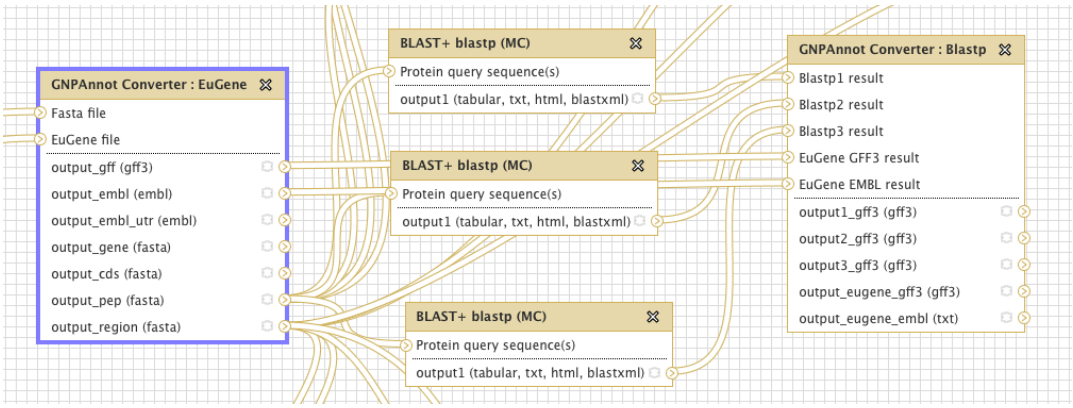

Figure 2 : Workflow Galaxy pour l'annotation fonctionnelle

Perfectionnement de l'annotation structurale :

Pour préciser la structure des gènes prédits (Figure 3), on utilise dans un premier temps une combinaison de TBLASTN et Exonerate sur les bases de données EST de riz (*Oryza sativa* et *Oryza glaberrima*) et de sorgho.

On utilise également en parallèle une combinaison de BLASTX/Exonerate et le programme Genome Threader, sur la séquence nucléique élargie entre gènes (Figure 4).

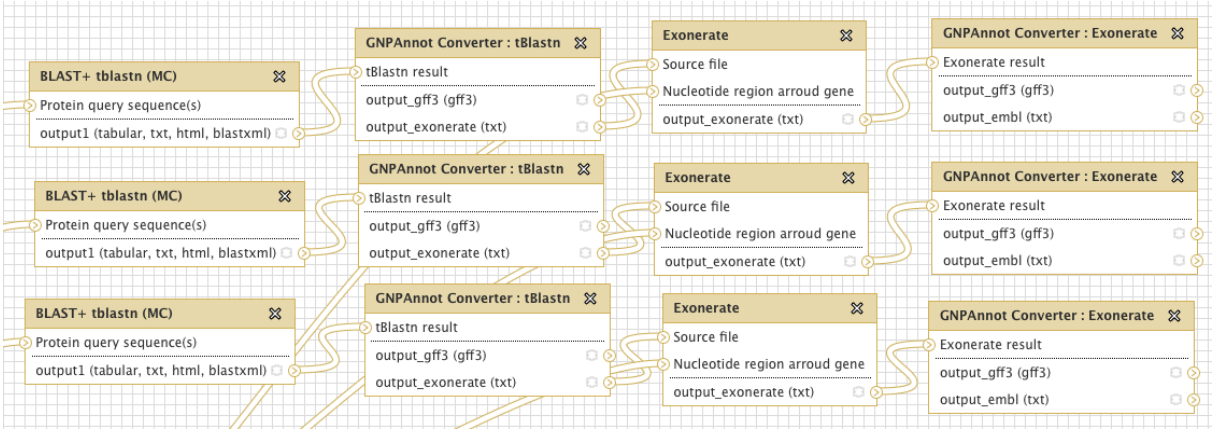

Figure 3 : Workflow Galaxy pour améliorer l'annotation structurale à partir des séquences protéiques des gènes prédits

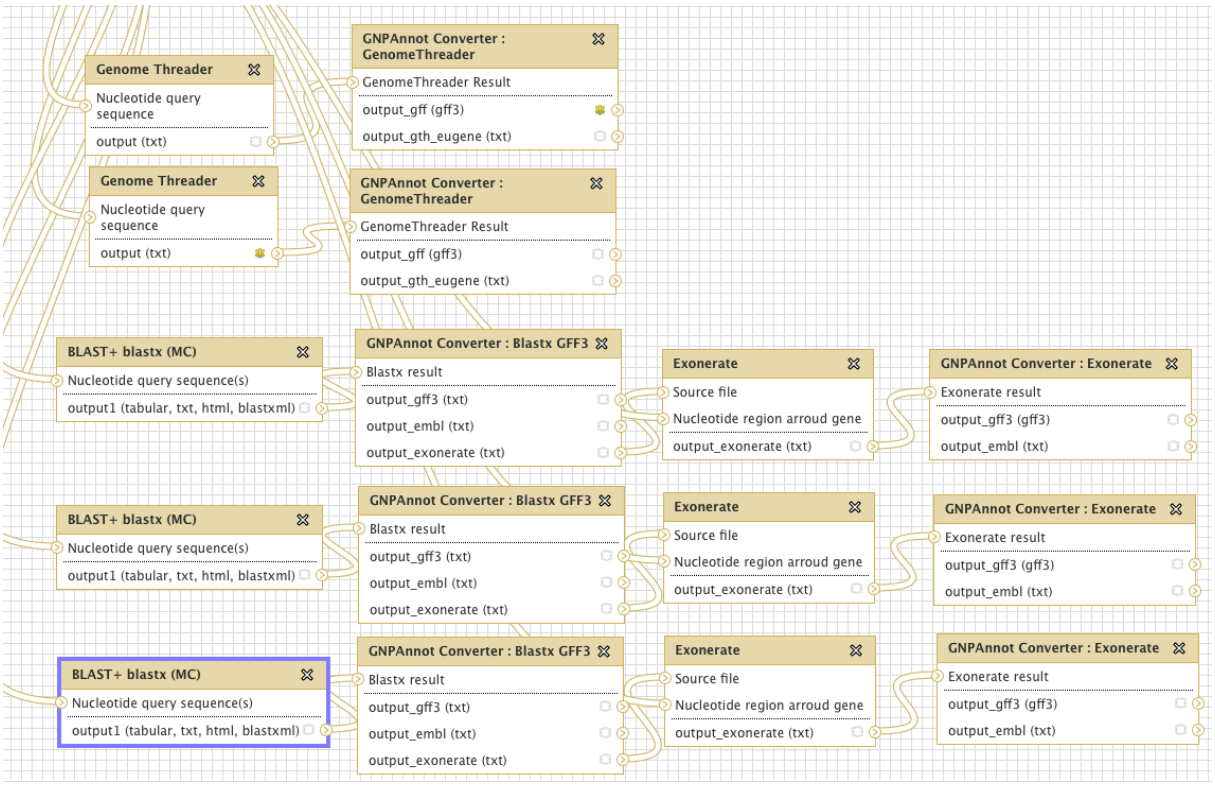

Figure 4 : Workflow Galaxy pour améliorer l'annotation structurale à partir des séquences nucléiques élargies des gènes.

*Récupération des fichiers de sortie du workflow:

Récupérez les fichiers de sortie suivants dans l'historique galaxy :

- (8) FGenesH (embl) : Fichier au format EMBL du logiciel FGenesH
- (44) EuGene (EMBL) : Fichier au format EMBL du programme EuGene
- (64) Exonerate OG_ngs (EMBL) : Fichier EMBL correspondant à la combinaison des programmes tBlastn/Exonerate sur les contigs de Riz (ssp. glaberrima)
- (66) Exonerate OS_mrnas (EMBL) : Fichier EMBL correspondant à la combinaison des programmes tBlastn/Exonerate sur la banque d'EST Riz (ssp japonica)
- (62) Exonerate SB_mrnas (EMBL) : Fichier EMBL correspondant à la combinaison des programmes tBlastn/Exonerate sur le banque d'EST sorgho.
- (72) Exonerate Rice (EMBL) : Fichier EMBL correspondant à la combinaison des programmes Blastx/Exonerate sur le protéome du Riz (MSU version 6.1)
- (70) Exonerate SwissProt (EMBL): Fichier EMBL correspondant à la combinaison des programmes Blastx/Exonerate sur la banque UniProtKB/SwissProt
- (68) Exonerate Sorghum (EMBL): Fichier EMBL correspondant à la combinaison des programmes Blastx/Exonerate sur le protéome du Sorgho
- (3) Os01_36429_36558.fna.repeat : Fichier au format BLAST tabulé correspondant à la comparaison de la séquences avec des bases de données de séquences répétées

3) Visualisation de l'annotation automatique avec « Artemis » et correction manuelle de l'annotation

→ Allez sur le site web de « Artemis » (sanger).
http://www.sanger.ac.uk/resources/software/artemis/

→ Cliquez sur l'onglet « download ».
→ Cliquez sur le bouton « launch » pour lancer le telechargement de l'application « Artemis »

Une fois « Artemis » téléchargé, l'application peut être exécutée en double-cliquant sur le fichier téléchargé. L'exécution d'Artemis nécessite au préalable l'installation de JAVA sur votre ordinateur (déjà réalisée pour le TD aujourd'hui).

Ouvrir le fichier d'annotation automatique « EuGene » au format EMBL dans artemis:

→ Cliquez sur le menu File/Read An Entry
Fichiers du type : Tous les fichiers
Nom de fichier: Nom de fichier: Galaxy___-[EuGene_(embl)].txt
A la question « there were warnings while reading - view now ? » répondez Non (si vous cliquez sur oui, ce n'est pas grave, mais vous aurez des avertissements (warnings) sur le format des annotations)

Q1: Quel est l'identifiant du premier gène prédit (= locus tag) ?

Q2 : Combien de structures codantes sont-elles prédites par Eugène ?

Création d'une piste (new entry) qui correspondra à la couche d'annotation manuelle:

→ Cliquez sur le menu Create/New Entry
Par défaut, votre nouvelle piste s'appelle « no name ». Pour renommer votre nouvelle piste :
→ Cliquez sur le menu Entries/Set Name Of Entry/no name
Une boite de dialogue s'ouvre. Tapez le nom que vous souhaitez donner à la piste. Par exemple « Manual_annotation ».
Copiez maintenant les features de la piste Galaxy___-[EuGene_(embl)].txt dans votre nouvelle piste. Pour être certain de ne copier que les features de cette entrée,
→ Vérifiez dans le menu Entries que seule la piste Galaxy___-[EuGene_(embl)].txt soit cochée. Dans le cas contraire décocher les autres pistes.
→ Dans la fenêtre où sont listées les features (celle contenant les information gene, CDS sous forme de texte),

sélectionnez toutes les features (cliquez dans la fenêtre et faire [Ctrl]+A)
→ Cliquez sur le menu Edit/Copy Selected Features To/Manual_annotation

Cette piste d'annotation sera utilisée pour réaliser les modifications. Elle constitue la piste d'annotation manuelle.

Ouvrir d'autres fichiers dans « Artemis » pour ajouter des couches d'informations supplémentaires:

Nom de fichier: Nom de fichier: Galaxy___-[FGenesH_(embl)].txt
Nom de fichier: Galaxy___-[Exonerate_OG_ngs_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_OS_mrnas_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_SB_mrnas_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_Rice_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_Sorgho_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_SwissProt_(EMBL)].txt
Nom de fichier: Galaxy___-[Os01_36429_36558.fna.repeat]
NB: Si vous voulez retirer une entrée : Menu Entries/Remove An Entry/choisissez le fichier à retirer

*Pour faciliter la visualisation des résultats :
→ Clic droit dans la fenêtre de visualisation des annotations
→ Cocher One Line Per Entry
→ Décocher Feature Labels

Comparaison de la structure prédite par Eugène et celle prédite par Fgenesh

→ Menu Entries/cochez la case correspondant à: Galaxy___-[FGenesH_(embl)].txt

Q3: Quelle est la différence majeure entre la prédiction EuGène et celle de Fgenesh ?

Q4: A quoi cela peut-il être dû ?
Aidez-vous de l'information contenue dans les différents fichiers que vous avez ouverts dans « Artemis » en cochant la case correspondant à la piste que vous souhaitez voir s'afficher (Menu Entries).

Utiliser les données de comparaison avec les bases de données de type banque EST/cDNA

→ Menu Entries/cochez la case correspondant à:
Galaxy___-[Exonerate_OS_mrnas_(EMBL)].txt (TBLASTN/Exonerate sur la banque d'EST Riz (ssp japonica)
Galaxy___-[Exonerate_SB_mrnas_(EMBL)].txt (TBLASTN/Exonerate sur le banque d'EST sorgho)

NB: Artemis est maintenant capable de lire les fichiers de mapping (fichiers de type bam) correspondant à l'alignement de données de type RNAseq sur la séquence génomique

Q5: Quelles sont les principales différences de structure entre la prédiction EuGène et celles reconstruites par « Exonerate »?

Q6: Au vu de ces informations, quelle structure vous parait la plus cohérente Fgenesh ou EuGene ?

Utiliser les données d'annotation d'espèces ou sous-espèces proches

→ Menu Entries/cochez la case correspondant à:
Galaxy___-[Exonerate_Sorghum_(EMBL)].txt (BLASTX/Exonerate contre protéome du sorgho)
Galaxy___-[Exonerate_OG_ngs_(EMBL)].txt TBLASTN/Exonerate sur les contigs de Riz (ssp. glaberrima)

Q7: A partir de l'ensemble des comparaisons, quelles sont les points de structure qu'il faut vérifier ? Quels types de données peuvent vous aider à prendre une décision ?

Question annexe sur la région du chromosome de riz analysée :
Q8: Est-ce que certaines informations vous permette d'établir une relation de micro-syntenie avec le Sorgho?

Utiliser les bases de données protéiques (Swissprot)

→ Menu Entries/cochez la case correspondant à:
Nom de fichier: Galaxy___ Galaxy___-[Exonerate_SwissProt_(EMBL)].txt (BLASTX/Exonerate contre base de données protéique UniprotKB/Swissprot

Q9: Est-ce que les résultats vous permettent de confirmer certains éléments de structure à corriger ?

Corriger la structure d'un gène

Les modifications seront réalisées sur les features de la piste d'annotation manuelle que vous avez créée.
Pour ajouter/modifier un exon dans un CDS
→ cliquez sur le CDS à corriger puis [Ctrl]+E
Une fenêtre d'édition s'ouvre. Les informations de coordonnées des exons sont listés dans la boite « 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
→ Ajoutez les coordonnées du nouvel exon puis validez en cliquant sur les boutons « Apply » puis « OK »

→ Vérifiez la jonction GT / AG des exons créés

Q10: Selon vous quelles sont les coordonnées correctes du premier exon du gène ?

→ Vérifiez à l'aide des informations dont vous disposez la structure du premier exon.
→ Modifiez et vérifiez les jonctions au niveau des sites d'épissage.

Q11: Selon vous quelles sont les coordonnées correctes du dernier exon du gène ?

→ Vérifiez à l'aide des informations dont vous disposez la structure du dernier exon.
→ Modifiez et vérifiez les jonctions au niveau des sites d'épissage.

Validation des corrections de l'annotation structurale

Afin de vérifier si les modifications de structure ont amélioré (ou pas) l'annotation du gène, vous allez comparer l'annotation initiale d'EuGene (Galaxy___-[EuGene_(embl)].txt) à votre annotation manuelle.
Pour cela nous allons aligner les protéines prédites par chacune des 2 annotations avec les protéines de la base de données Uniprot/Trembl.

Récupérez la séquence protéique du premier gène sur la piste EuGene et sur la piste annoté manuellement
→Clic droit sur l'objet CDS
→View/Amino acids of selection as fasta
Copier/coller la séquence dans un éditeur de texte (bloc note) et enregistrer chacune des séquences.
Par exemple : locus_tag_ori.faa (pour la séquence EuGene) et locus_tag_cor.faa (pour votre annotation corrigée).

→ Lancez un navigateur, ouvrez deux onglets et aller à l'adresse suivante
http://www.expasy.ch/tools/blast/
ou http://www.uniprot.org/ onglet Blast
→ Copier-coller la séquence locus_tag_ori.faa dans un des onglets et celle de locus_tag_cor.faa dans l'autre.
Lancer le BLASTp en cliquant sur le bouton Run BLAST

Q12 : Observez les alignements, votre annotation est-elle meilleure que l'annotation automatique initiale ?
Quels indices vous permettent de conclure ?
Quelle modification supplémentaire sur la structure du gène pourriez-vous apporter ?
→ Affinez la structure de votre gène et vérifiez les bornes aux sites d'épissage.

Annotation fonctionnelle du gène

Lancez « InterproScan » pour la recherche de domaines protéiques conservés et tenter d'identifier la famille et la fonction de votre protéine
Lancez un navigateur internet et aller sur le site : http://www.ebi.ac.uk/Tools/pfa/iprscan/
Récupérez la séquence protéique du premier gène de votre piste d'annotation manuelle
→Clic droit sur l'objet CDS
→View/Amino acids of selection as fasta
→Copier et coller la séquence dans la fenêtre de « InterproScan »
→Lancez la recherche Interpro.

→ Lancez un navigateur et aller à l'adresse suivante : http://www.expasy.ch/tools/blast/ ou http://www.uniprot.org/ onglet Blast
→ Copier-coller la séquence protéique du premier gène de votre piste d'annotation manuelle.
Lancer le BLASTP en cliquant sur le bouton Run BLAST

Analysez vos alignements blastp contre Uniprot et le résulat d' »InterproScan »

Q13: Quelle est l'accession Uniprot correspondant à votre gène ?
Q14: Quels sont les domaines conservés qu' « InterproScan » a détecté.
Q15: Grâce à ces éléments attribuez la fonction putative de votre gène en utilisant un vocabulaire contrôlé.

→Cliquez sur le CDS puis [Ctrl]+E pour ouvrir la boite d'édition du gène à annoter.
→Vérifiez que la fonction putative du gène soit bien renseignée dans le qualifier « product »
Ajoutez des qualifiers et les renseigner:
/evidence=curated
/status="finished"

Annotation des éléments répétés

Au cours du TD, vous avez repéré la présence d'éléments répétés dans un intron du gène.
Pour délimiter la région répétée et tenter de définir le type d'élément présent vous allez récupérer la séquence de l'intron et faire une recherche de similarité contre une base de données d'éléments répétés de riz

→ Notez les coordonnées de l'intron (début et fin)
→Créez une nouvelle feature : Menu Create/New Feature
→Par défaut la clé « Key » est renseignée « misc_feature ». Changez en « repeat_region »
→Entrez les coordonnées de l'intron dans la fenêtre « location » : 3019..6125 et valider (« apply » puis « OK »)
→Clic droit sur l'objet repeat_region
→View/bases of selection as fasta
→Copier et coller la séquence dans la fenêtre de « Censor » à l'adreese suivante : http://www.girinst.org/censor/
→Selectionnez la base de TEs riz dans la liste de choix
→ Lancez la recherche
→ Notez les coordonnées et le type d'éléments dectéctés.
→ Modifiez votre annotation en conséquence en éditant votre feature « repeat_region »

- Coordonnées
- Informations dans un qualifier /note="ma famille de TE identifiée"
- /annotator_comment="renseignez les informations que vous souhaitez conserver, par exemple la manière dont vous avez procédé pour identifier le TE"

Sauvegarde de votre annotation manuelle

Pour enregistrer votre annotation manuelle :

→Menu File/Save An Entry As/EMBL format/Ma_piste_manuelle
Donnez un nom au fichier et sauvergardez.

→Fermez artemis.

Q16: A partir des notions que vous avez acquises lors de ce TD, pouvez-vous visualiser l'annotation automatique et manuelle de la région analysée ?