Détection de Gènes
Présentation d'un outil : Funannotate predict

COQUERELLE.M SANCHEZ-MOREIRA.H

LaTeX

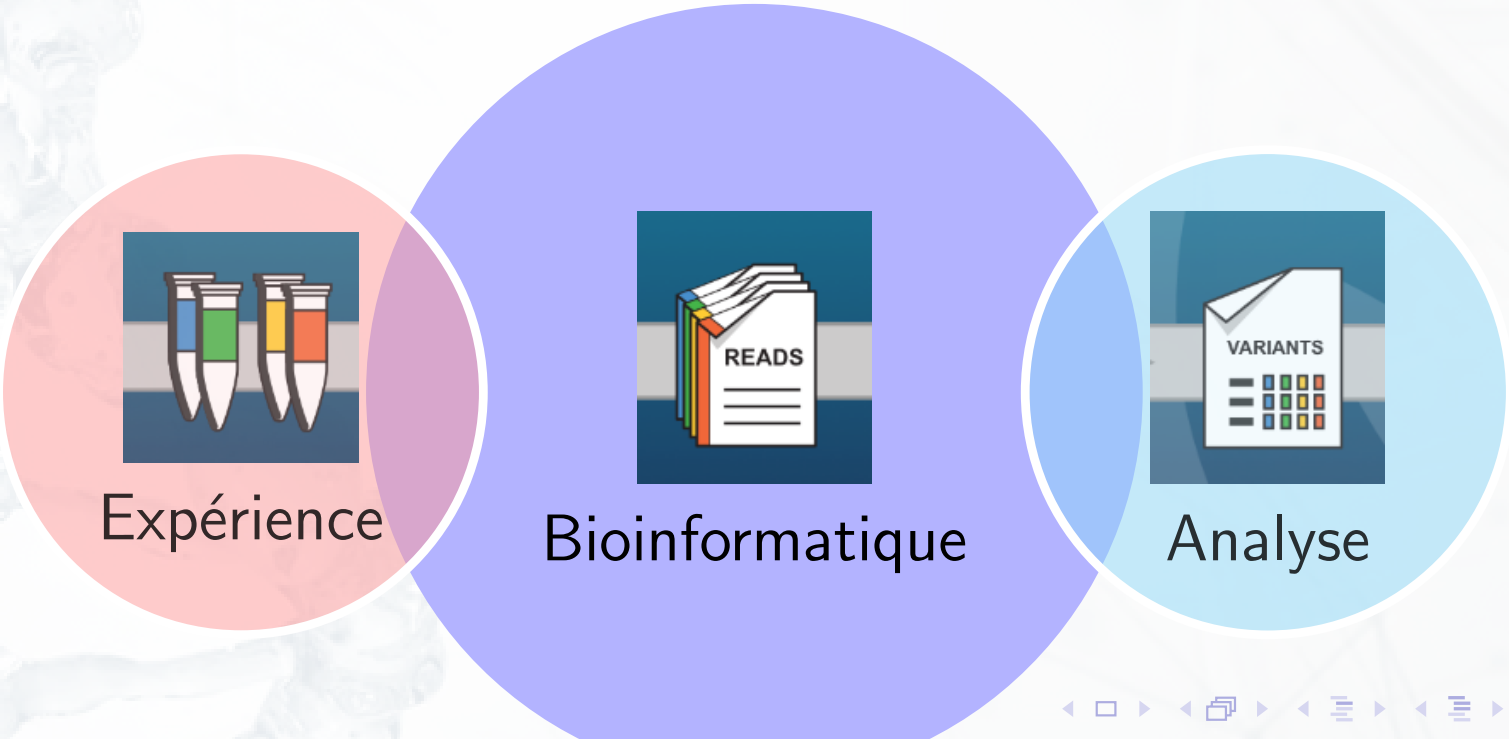
October 7, 2024

Sommaire

Processus d'identification de gène

Comment définir cette étape ?

Combinaison des étapes expérimentales, bioinformatiques et d'expertise biologique qui aboutissent à l'identification d'un ou des gène(s) pour une séquence nucléotidique donnée.



Objectifs

Comprendre les fonctions biologiques, diagnostiquer des maladies génétiques, développer des thérapies ciblées.

Contraintes

- Complexité/diversité des génomes et des variations génétiques.
- Coût : Technologie et analyse bioinformatique coûteuses.
- Temps : Processus de séquençage et d'analyse long.
- Précision : Nécessité d'une haute précision pour éviter les faux positifs/négatifs.

Quelques mots sur la phase expérimentale ...

Méthodes conventionnelles

- Caryotype : Analyse des chromosomes pour détecter des anomalies structurelles.
- Séquencage de Sanger : Détection de variations génétiques spécifiques.

Actuel et futur

- Séquençage NGS : parallélisation.
- Séquençage 3GS : Nanopore...

Quelques mots sur la diversité des solutions ...

Solutions bioinformatiques

- Diversité de solutions : BLAST, GATK, ANNOVAR.
- Algorithmique : Alignement de séquences, détection de variants.
- Présentation d'un outil particulier : funannotate predict[1]

Installation via environnement python

Option 1 :

- Portabilité.
- Isolation de l'environnement de travail.
- Installation pas très intuitive.

```
# Environnement virtuel python
~$ python3 -m venv env-fun
~$ source env-fun/bin/activate

# Installation de funannotate
~$ python3 -m pip install funannotate
```

Installation via Docker

Option 2 :

- Simplification de l'installation.
- Portabilité.
- Isolation de l'environnement de travail.

```
# Installer fun annotate via docker :
~$ sudo snap install docker ;
~$ sudo docker pull nextgenusfs/funnanotate
~$ wget -O funannotate-docker https://LienGITHUB/

~$ chmod +x funannotate-docker
~$ sudo mv funannotate-docker /usr/local/bin
```

Présentation générale : entrées

Contexte

- Un outils historiquement conçu pour l'étude fongique.
- Généralisé ensuite à toutes les espèces.

Les entrées : Expérience unitaire ou banque de données

- Fichier FASTA du génome
- Transcrits Trinity
- Fichier BAM des alignements RNAseq
- Données PASA/TransDecoder
- Transcrits EST

Présentation générale : processus

Prédicteurs de gène utilisés et fonctionnalités [1]

- Alignement des évidences de transcrits / protéiques sur une référence: `minimap2, Diamond/Exonerate`
- Prédiction de gènes *ab initio* : `GeneMark-ES/ET, Augustus, snap, GlimmerHMM, CodingQuarry`
- Combinaison de prédictions pour générer des modèles consensuels :`Evidence Modeler (EVM)`

Présentation générale : sorties

Fichiers de sortie

- Table d'annotation NCBI : `genome.tbl`
- Fichier GenBank : `genome.gbk`
- FASTA des protéines codantes : `proteins.fa`
- FASTA des transcrits :`transcripts.fa`

Test unitaire : Funannotate en action !

Détection de gènes sur un génome fongique

- On télécharge une séquence génomique sur NCBI.
- Préparation de la séquence d'intérêt (masquage RET)

```
# Fichier de sortie et configuration :
~$ touch Output.fasta
~$ unzip Aspergillus_nidulans.zip
~$ sudo funannotate-docker mask -i
   ../GCF_000011425.1_ASM1142v1_genomic.fna
   -o output.fasta
```

Test unitaire : Funannotate en action !

Périmètre de l'analyse, cas des éléments répétés

- L'évaluation se peut se faire en précisant des modèles de prédictions (ici Augustus), avec une référence associée (ici Aspergillus Nidulans).

Prétraitement :[2]

```
# Formatage des en-tetes :
~$ sed -i 's/Aspergillus nidulans FGSC A4
    chromosome//g' Output.fasta

# Demarrage de l'analyse :
~$ sudo funannotate-docker predict
     -i Output.fasta
     -o ~/Rep/Dos -s 'Aspergillus nidulans'
     --augustus_species anidulans
```

```
mickael@mickael-Inspiron-16-5630:~/Bureau/Nouveau_dossier$ sed -i 's/Aspergillus nidulans FGSC A4 chromosome //g' Output.fasta
mickael@mickael-Inspiron-16-5630:~/Bureau/Nouveau_dossier$ sudo funannotate-docker predict -i Output.fasta -o ~/Bureau/Nouveau_dossi
-------------------------------------------------------
[Oct 02 09:13 AM]: OS: Debian GNU/Linux 10, 16 cores, ~ 16 GB RAM. Python: 3.8.12
[Oct 02 09:13 AM]: Running funannotate v1.8.17
[Oct 02 09:13 AM]: GeneMark not found and $GENEMARK_PATH environmental variable missing. Will skip GeneMark ab-initio prediction.
[Oct 02 09:13 AM]: Skipping CodingQuarry as no --rna_bam passed
[Oct 02 09:13 AM]: Parsed training data, run ab-initio gene predictors as follows:
  Program      Training-Method
  augustus     pretrained
  glimmerhmm   busco
  snap         busco
[Oct 02 09:13 AM]: Loading genome assembly and parsing soft-masked repetitive sequences
[Oct 02 09:13 AM]: Genome loaded: 8 scaffolds; 29,828,291 bp; 6.64% repeats masked
/venv/lib/python3.8/site-packages/funannotate/aux_scripts/funannotate-p2g.py:14: DeprecationWarning: pkg_resources is deprecated as a
  from pkg_resources import parse_version
[Oct 02 09:13 AM]: Mapping 558,971 proteins to genome using diamond and exonerate
[Oct 02 09:18 AM]: Found 277,295 preliminary alignments with diamond in 0:04:53 --> generated FASTA files for exonerate in 0:00:14
    Progress: 277295 complete, 0 failed, 0 remaining
```

Analyse séquentielle : 2 heures plus tard

- Gènemark absent et "CodingQuarry " ignoré (.bam)
- Liste prédicteurs et méthodes d'entrainements utilisés
- Diamond et Exonerate : 559 000 protéines mappées
- 277 295 alignements préliminaires de Diamond (Gen.fasta)

```
[Oct 02 11:31 AM]: Exonerate finished in 1:33:33: found 2,976 alignments
[Oct 02 11:31 AM]: Running BUSCO to find conserved gene models for training ab-initio predictors
[Oct 02 11:54 AM]: 1,280 valid BUSCO predictions found, validating protein sequences
[Oct 02 11:55 AM]: 1,278 BUSCO predictions validated
[Oct 02 11:55 AM]: Running Augustus gene prediction using anidulans parameters
    Progress: 63 complete, 0 failed, 0 remaining
[Oct 02 12:09 PM]: 8,193 predictions from Augustus
[Oct 02 12:09 PM]: Pulling out high quality Augustus predictions
[Oct 02 12:09 PM]: Found 386 high quality predictions from Augustus (>90% exon evidence)
[Oct 02 12:09 PM]: Running SNAP gene prediction, using training data: /home/mickael/Bureau/Nouveau_dossier/predict_misc/busco.final.gff3
[Oct 02 12:10 PM]: 0 predictions from SNAP
[Oct 02 12:10 PM]: SNAP prediction failed, moving on without result
[Oct 02 12:10 PM]: Running GlimmerHMM gene prediction, using training data: /home/mickael/Bureau/Nouveau_dossier/predict_misc/busco.final.gff3
[Oct 02 12:14 PM]: 10,884 predictions from GlimmerHMM
[Oct 02 12:14 PM]: Summary of gene models passed to EVM (weights):
  Source          Weight   Count
  Augustus        1        8177
  Augustus HiQ    2        393
  GlimmerHMM      1        10884
  Total           -        19454
[Oct 02 12:14 PM]: EVM: partitioning input to ~ 35 genes per partition using min 1500 bp interval
    Progress: 293 complete, 0 failed, 0 remaining
```

3 heures plus tard ...

- 2976 alignements conservés avec Exonerate
- 1278 prédictions de genes de BUSCO
- 8197 prédictions de AUGUSTUS avec le paramètre anidulans[3]
- 386 prédictions de haute qualité de AUGUSTUS
- Pas de prédiction SNAP, 10 884 prédictions de GlimmerHMM
- Liste des prédictions et des poids,démarrage de Evidence Modeler (EVM)

```
[Oct 02 12:28 PM]: Converting to GFF3 and collecting all EVM results
[Oct 02 12:28 PM]: 10,510 total gene models from EVM
[Oct 02 12:28 PM]: Generating protein fasta files from 10,510 EVM models
[Oct 02 12:28 PM]: now filtering out bad gene models (< 50 aa in length, transposable elements, etc).
[Oct 02 12:28 PM]: Found 184 gene models to remove: 2 too short; 15 span gaps; 167 transposable elements
[Oct 02 12:28 PM]: 10,326 gene models remaining
[Oct 02 12:28 PM]: Predicting tRNAs
[Oct 02 12:29 PM]: 175 tRNAscan models are valid (non-overlapping)
[Oct 02 12:29 PM]: Generating GenBank tbl annotation file
[Oct 02 12:29 PM]: Collecting final annotation files for 10,501 total gene models
[Oct 02 12:29 PM]: Converting to final Genbank format
[Oct 02 12:29 PM]: Funannotate predict is finished, output files are in the /home/mickael/Bureau/Nouveau_dossier/predict_results folder
[Oct 02 12:29 PM]: Your next step might be functional annotation, suggested commands:
-------------------------------------------------------
Run InterProScan (manual install):
funannotate iprscan -i /home/mickael/Bureau/Nouveau_dossier -c 2

Run antiSMASH (optional):
funannotate remote -i /home/mickael/Bureau/Nouveau_dossier -m antismash -e youremail@server.edu

Annotate Genome:
funannotate annotate -i /home/mickael/Bureau/Nouveau_dossier --cpus 2 --sbt yourSBTfile.txt
-------------------------------------------------------

[Oct 02 12:29 PM]: Training parameters file saved: /home/mickael/Bureau/Nouveau_dossier/predict_results/anidulans.parameters.json
[Oct 02 12:29 PM]: Add species parameters to database:

  funannotate species -s anidulans -a /home/mickael/Bureau/Nouveau_dossier/predict_results/anidulans.parameters.json
```

5 heures plus tard ....

- 10 510 modèles de gènes par EVM.
- 10 326 après filtrage des "mauvais modèles"
- 175 d'ARNt
- Création des fichiers résultats pour les 10 501 modèles de gènes.

"Aspergillus nidulans" AND "genome" AND "assembly" [4]

	RefSeq	GenBank
Provider	Eurofungbase (Eurofung)	Eurofungbase (Eurofung)
Name	Annotation submitted by Eurofungbase (Eurofung)	Annotation submitted by Eurofungbase (Eurofung)
Date	Apr 13, 2023	Mar 1, 2015
Genes	10 518	10 597
Protein-coding	10 455	10 534
	View RefSeq annotation	View GenBank annotation

$$\Delta \approx 0,07\%$$

Conclusion

- Un outil puissant pour détecter les gènes, facilitant l'annotation génomique.
- Son utilisation en environnement virtuel ou via Docker offre de la flexibilité et isolation dans son utilisation
- Précision et efficacité notables dans l'annotation de génomes fongiques.
- Des améliorations continues et des mises à jour.
- Pour aller plus loin : applications potentielles sur d'autres espèces avec un matériel adapté pour confirmer sa fiabilité.

[1] J Palmer. *Funannotate*. original-date: 2015-12-18T20:21:15Z. Aug. 2023. URL: https://github.com/nextgenusfs/funannotate (visited on 10/04/2024).

[2] A Fiston-Lavier. "Introduction à Linux et lignes de commandes". fr. In: (Sept. 2024). URL: https://moodle.umontpellier.fr/pluginfile.php/1524190/mod_resource/content/0/CM_IntroLinux.pdf (visited on 09/27/2024).

[3] M Stanke. *The AUGUSTUS gene prediction tool*. en. Text.Homepage.Organizational. Publisher: Institute for Mathematics and Computer Science, University of Greifswald. June 2003. URL: https://bioinf.uni-greifswald.de/augustus/.

[4] Conrad L Schoch et al. "NCBI Taxonomy: a comprehensive update on curation, resources and tools". en. In: *Database* 2020 (Jan. 2020), baaa062. ISSN: 1758-0463. DOI:

MERCI DE VOTRE ATTENTION !