Compte Rendu Visio Conférence *PepiAnnot*
18 avril 2024

Lors de la visio PepiAnnot :
- Il y avait 23 personnes connectées
- Toutes les infos sur le site du PEPI IBIS https://pepi-ibis.inrae.fr/annotation-genomes

Ordre du jour :

---

1. Futurs thèmes à aborder

- Annotation des familles de gènes : tester Helixer dans des familles de gènes très conservées
- Pan-génome : action des PF Bioinfo INRAE à l'automne 2024
- Véracité des extrémités des gènes (Véronique) : par exemple pb d'annotation entre TAIR10 et ARAPORT11
- Voir les autres thèmes proposés dans les CR précédents

2. Coanimation de PEPI-Annot avec GT Annot de MERIT

- MERIT : réseau métiers du CNRS, objectif équivalent au PEPI IBIS pour rassembler les bioinformaticien autour de thématiques communes. Rassembler les bioinfo afin d'éviter les personnes isolées.
- Groupe Annotation : Sophie Lemoine, Erwan Corre
- Animer un groupe commun avec le PEPI-Annot autour de l'annotation des génomes

3. Retour d'expérience sur Helixer par Alexandre Cormier (Ifremer)

Enregistrement vidéo sur le site

Helixer : outil d'annotation des génomes Eucaryotes, basé sur de l'IA

Constat : encore 80% des génomes séquencés non annotés donc besoin d'un outil rapide

Possible maintenant car se base sur l'ensemble des génomes annotés et utilisation des ressources informatiques puissantes (GPU).

Ligne de commande très simple : un génome format fasta, chromosomes ou scaffolds, 4 modèles possibles (champignons, plantes, invertébrés, vertébrés)

Principe :

- Définit des classes (UTR, codant, introns, intergénique etc.) et ensuite associe chaque région à un classe (modèle HMM). Deeplearning CNN
- Temps de calculs bcp plus rapide que les autres, plutôt quelques heures que quelques jours
- Pas de pré-processing : pas de masquage des répétitions, par contre n'annote pas les éléments transposables.
- Besoin de GPU pour la partie Deeplearning, images singularity et Docker
- Entrainement fait sur une dizaine de génomes et validé sur plus de 200 génomes.
- Sortie : GFF3 qui est correctement écrit !

Validation, métriques
AGAT, BUSCO (completness) et un autre outil OMArk plus complet que BUSCO par rapport au protéome généré

Jeux de données : huitres, plantes, insectes
Comparaison avec Braker3

Quelques résultats :
Dans la plupart des cas Helixer prédit plus de gènes (mRNA=1 gène), valider en véifiant via outils de validation et remapping sur génomes avec nouvel annotato
Avantages :

- Très rapide, pas besoin de masquer les repeat
- Très simple d'utilisation, très peu de paramètres
- Pas besoins de RNA-seq au contraire de la plupart des outils d'annotation
- Permet d'obtenir une première annotation très fiable
- Semble marcher sur des familles de gènes conservés → mais à vérifier avec d'autres familles
- Prédit les UTR des gènes, chaque mRNA prédit correspond à un locus=1 gènes

Limites :

- Sur un très gros génome polyploïde Helixer semble perdu, retour de Damien Hinsinger
- Ne prédit que les gènes codants pour des protéines
- Pas d'isoformes prédits : 1mRNA=1gènes
- Sur un génome a donné des UTRs bcp plus petits que ceux prédits par Braker3, donc peut-etre un pb de fiabilité des extrémités
- Met des exons d'1bp, aucun paramètre sur les modèles de gènes (taille des introns, exons etc.)