Projet Annotation – Description des parsers

Parser pour les résultats de GeneMark :

Les résultats de GeneMark qui se présentent de la façon suivante :

GeneMark permet de prédire les régions codant pour des protéines sur le génome (CDS). Quand DNA strand est indiqué direct, left end correspond au début prédit et right end à la fin.  Si DNA strand est indiqué complément c’est l’inverse. Dans la prédiction, il n’y a pas d’ambiguieté pour prédire la fin (codon stop) mais il peut y avoir plusieurs options possibles pour le début et c’est en combinant différentes approches que l’on tente de le définir correctement. Pour vous entrainer l’idée est de garder toutes les prédictions de GeneMark sans éditer votre fichier GeneMark avant pour ne garder que le résultat qui vous semble correct. Ceci est pour vous entraîner à créer des compteurs (cf ci-dessous).

GENEMARK PREDICTIONS

Sequence: BS09819

Sequence file: seq.fna

Sequence length: 14616

GC Content: 37.49%

Window length: 96

Window step: 12

Threshold value: 0.400

---

Matrix: Bacillus\_subtilis\_168

Matrix author: -

Matrix order: 4

List of Open reading frames predicted as CDSs, shown with alternate starts

(regions from start to stop codon w/ coding function >0.40)

Left Right DNA Coding Avg Start

end end Strand Frame Prob Prob

-------- -------- ---------- ----- ---- ----

1 276 direct fr 1 0.61 ....

28 276 direct fr 1 0.61 ....

112 276 direct fr 1 0.47 0.00

495 3587 direct fr 3 0.41 0.12

528 3587 direct fr 3 0.41 0.07

1191 3587 direct fr 3 0.48 0.65

1356 3587 direct fr 3 0.45 0.00

1467 3587 direct fr 3 0.44 0.07

12559 12906 complement fr 3 0.66 0.03

12559 12894 complement fr 3 0.68 0.13

12559 12885 complement fr 3 0.71 0.29

12559 12870 complement fr 3 0.73 0.32

12559 12699 complement fr 3 0.58 0.01

12559 12696 complement fr 3 0.54 0.01

Le format du fichier gff comprend :

Nom\_seq Méthode-prédiction feature début fin score brin frame note

Pour brin + = direct

- = complementent

Pour frame, on mettra un ‘. ‘ qui indique que l’on ne remplit pas le champ

Dans note, on met les commentaires que l’on veut. Donc là je propose que vous intégriez comme ci\_dessous le numéro de la CDS et de sa version (une version par codon start proposé dans la sortie GeneMark). Vous pouvez associer une deuxième annotation qui est la position du codon start (feature ATG ici) déduite des résultats de la prédiction.

Les résultats ci-dessus sont à reformeter comme suit :

BS09819 GenMark CDS 1 276 0.68 + . gene GM\_CDS\_1.1

BS09819 GenMark CDS 28 276 0.68 + . gene GM\_CDS\_1.2

BS09819 GenMark CDS 112 276 0.58 + . gene GM\_CDS\_1.3

BS09819 GenMark CDS 1191 3587 0.51 + . gene GM\_CDS\_2.1

BS09819 GenMark ATG 1191 1193 0.32 + .

BS09819 GenMark CDS 2397 3587 0.51 + . gene GM\_CDS\_2.2

BS09819 GenMark ATG 2397 2399 0.51 + .

BS09819 GenMark CDS 2589 3587 0.59 + . gene GM\_CDS\_2.3

BS09819 GenMark ATG 2589 2590 0.97 + .

BS09819 GenMark CDS 2619 3587 0.57 + . gene GM\_CDS\_2.3

BS09819 GenMark ATG 2619 2621 0.02 + .

...

Le parser pour GeneMark.hmm est plus simple car il n’y a pas de début alternatif proposé.

Le parser sacn\_for\_matches :

La prédiction de motif particulier sur la séquence (RBS, promoteur et terminateur de transcription) obtenue en utilisant le programme scan\_for\_matches concernant la prédiction des RBS et donc du codon initiateur, des promoteurs et des terminateurs de transcription. Un seul programme doit pouvoir permettre de parser les différents résultats de scan\_for\_matches.

Précision pour le parser des résultats de scan\_for\_matches

Faire un seul programme pour parser les résultats des trois features (RBS, promoteur, terminateur). Pour cela passer en argument le nom de fichier d'entrée, la feature, le nom du fichier de sortie.

Il faut transformer les résultats qui sont du type :

>BS09819:[347,378]

ttgata tttttttgatttttagaatg tatagt

>BS09819:[1038,1084]

ttgata aatgaactgtgtggtgaatctgctgcatatcaaga tataat

pour créer le fichier GFF ayant le format suivant (chaque colonne est séparée de la suivante par une tabulation) qui comme pour GeneMark doit contenir les infos par colonne correspondant à (pas de valeur pour le score et le frame donc des . à la place:

nomseq method feature posgauche posdroite score strand frame note motif

Donc ici :

BS09819 Patscan promoter 347 378 . + . note "ttgata tttttttgatttttagaatg tatagt"

BS09819 Patscan promoter 1038 1084 . + . note " ttgata atgaactgtgtggtgaatctgctgcatatcaaga tataat"

Pour prédiction RBS feature = RBS etc.