1. TP: annotation de conservation

1.1 Description du projet

Dans cet exercice, on produit des pistes d'annotation sur conservation dans un génome. On se sert des alignements de régions orthologues dans génomes du groupe *Saccharomyces sensus stricto*. Six génomes de ce groupe ont des séquences de haute qualité [Scannell, D. R. et al. "The awesome power of yeast evolutionary genetics : New genome sequences and strain resources for the Saccharomyces sensu stricto genus," *G3* **1** :11 (2011)]. Vous travaillez avec données du site http://www.saccharomycessensustricto.org/ pour cinq espèces : *S. cerevisiæ* (Scer), *S. paradoxus* (Spar), *S. mikataæ* (Smik), *S. kudriavzevii* IFO 1802^{T} (Skud), et *S. bayanus var. uvarum* CBS 7001 (Sbay). Vous devez créer des annotations du génome de *Saccharomyces cerevisiæ* dans le fureteur UCSC Genome Browser (http://genome.ucsc.edu/, Genome Browser → `Other` (group)/`S.cerevisiae` (genome)/`sacCer3` (assembly)). Afin de produire les annotations, on se sert des alignements par le logiciel BigFoot [Satija, R. et al. "BigFoot : Bayesian alignment and phylogenetic footprinting with MCMC," *BMC Evolutionary Biology*, **9** :217 (2009)].

1.2 Tâches (20 points)

a. Rencontre avec BigFoot (1 point). BigFoot est un logiciel d'alignement statistique basé sur un modèle de substitutions et insertions/suppressions (indels) au long d'un arbre phylogénétique. La méthode utilise un modèle de Markov caché avec deux états (régions d'évolution lente et rapide) qui ont leur propres taux de substitution et d'indel. (Régions conservées ont peu de trous et peu de substitutions.) Téléchargez le logiciel [http://sourceforge.net/projects/bigfoot/], lisez l'article [Satija et al (2009)], et étudiez la documentation [http://www.stats.ox.ac.uk/~satija/BigFoot/doc/help/].

b. Utilisation de BigFoot (4 points). Choisir un alignement multiple de régions intergéniques orthologues à http://www.saccharomycessensustricto.org/ (menu Intergenics →). Vous avez le droit de choisir n'importe quel alignement mais vérifiez qu'il est au moins 200 pb de longueur et qu'il contient tous les cinq espèces Scer, Spar, Smik, Skud, et Sbay. Utiliser BigFoot avec cet alignement et la phylogénie spécifiée sur le site du cours IFT6299. Faire :

1. Changer le nom des séquences dans le fichier Fasta à Scer, Spar, ... (il faut utiliser le même nom que l'arbre). Mettre Scer au premier (BigFoot prend le premier séquence comme référence à annoter).
2. Lancer BigFoot (avec mémoire assez grand spécifié : `java -Xmx4096M -jar BigFoot.jar`).
3. Charger l'alignement et l'arbre (menu File), cocher tout pour la sortie (File → Preferences), sélectionner modèle HKY85 (Model → Hasegawa/Kishino/Yano 1985), mettre *burn-in* à au moins 50000 (MCMC → Settings). Presser MCMC → Run.

L'alignement statistique de BigFoot est basé sur une approche MCMC (*Markov Chain Monte Carlo*) : il génère des **échantillons** (*samples*) de paramètres d'alignement, séquences d'états lentes/rapides, alignements de séquences d'ADN. En une exécution (*run*), le logiciel produit des échantillons randomisés en $n_0 + n$ cycles, où n_0 est le temps «burn-in» qui est le nombre d'échantillons à ignorer au début (pendant que l'échantillonnage randomisé s'approche de l'optimum, ici $n_0 = 50000$). Après le «burn-in», on enregistre les solutions échantillonnées pendant n cycles. Avec un taux d'échantillonnage (*sampling rate*) 2500, la sortie comprend les échantillons aux cycles $n_0 + 2500$, $n0 + 5000$, $n_0 + 7500$, ..., $n_0 + n$. bigFoot écrit la sortie dans de fichiers dont le nom vient du nom de fichier d'alignement. Exemple : dans le cas du fichier

d'alignement `YJL178Cu_YJL177Wu.mfa` (format Fasta), BigFoot écrit les fichiers `YJL178Cu_YJL177Wu.log` (alignements à chaque échantillon MCMC), `YJL178Cu_YJL177Wu.ll` (log-vraisemblance de l'échantillon), `YJL178Cu_YJL177Wu.mpd` (alignement MPD), `YJL178Cu_YJL177Wu.pred` (probabilité d'état lent de chaque position dans la séquence de référence). On produit les annotations à partir des fichiers `pred` et `mpd`.

c. Produire une piste d'annotation (5 points). Transformez la sortie de BigFoot en pistes d'annotation de format WIG [http://genome.ucsc.edu/goldenPath/help/wiggle.html]. Le fichier `.pred` donne la probabilité postérieure de l'état de conservation pour chaque résidue de la première séquence de l'alignement. Mettez les probabilités dans un fichier WIG. Afin d'obtenir les coordonnées génomiques, examinez coordonnées pour les gènes à coté de votre région choisie dans le fureteur *Yeast Genome Browser* (http://www.yeastgenome.org), ou les fichiers d'annotation GFF (http://www.saccharomycessensustricto.org/, Annotations →). Créer un deuxième fichier d'annotation WIG à partir des probabilités de colonnes dans la sortie `.mpd`. (Attention : ignorer les colonnes avec trous dans Scer). Vérifiez vos annotations dans le UCSC Genome Browser (Manage custom tracks →).

d. Obtenir des séquences ADN (5 points). On veut produire une piste d'annotation comme en c. pour un intron. Pour cela, il faut extraire les séquences orthologoues. Tout gène de levure possède un identifieur SGD[1] (p.e., YJL177W) qui correspond à des identifieurs spécifiques dans les génomes à http://www.saccharomycessensustricto.org (p.e., Scer_10.56 dans Scer, Sbay_6.258 dans Sbay, etc.). Vous trouvez un tableau d'orthologues à http://www.saccharomycessensustricto.org/SaccharomycesSensuStrictoResources//tables/Supp_Table1_OrthologSets.xls. Téléchargez les fichiers de http://www.saccharomycessensustricto.org/ :

- ⋆ génomes assemblés (Assemblies → [Unordered/All] Scaffolds) : `Scer.scaffolds`, ...
- ⋆ annotations (Annotations → GFF) de format GFF [http://www.sanger.ac.uk/resources/software/gff/spec.html] : `Scer.gff`, ...
- ⋆ protéines (Annotations → Proteins) : `Scer.aa`, ...

Choisir un des gènes sur la liste suivante : YJL177W, YKL180W, YGL251C, YNL096C, YOR096W, YIL133C, YOL120C. Le gène contient un intron. Afin d'obtenir sa séquence génomique dans tous les génomes, on va aligner la séquence protéique du gène avec la région génomique annotée (hélas, l'annotation GFF ne montre pas les introns ici).

1. Vérifier les identifieurs spécifiques pour le gène choisi dans le tableau d'orthologues. Chercher les coordonnées génomiques dans les fichiers GFF. (P.e., gène YJL177W dans SGD est le gène Scer_10.56 (Gene attribute) dans `Scer.gff`, avec coordonnées chr10 :90786..91657.)
2. Sauvegarder la séquence du gène (toute la région) dans un fichier Fasta : utilisez un outil existant, ou implantez votre propre logiciel pour choisir une région par coordonnées à partir d'un fichier Fasta à l'entrée : getSequence`Scer.scaffolds`Scer_10 90786..91657 devrait écrire la sortie Fasta

   ```
   >Scer_10 /substring=90786..91657
   ATGGCAAGAT...
   ```

   Notez que la sortie doit être le complément inverse de la région génomique quand le gène se trouve sur le brin opposé — vérifiez l'orientation dans le fichier GFF.
3. Aligner la séquence protéique du gène avec la région génomique. Utilisez le logiciel exonerate [Slater et Birney "Automated generation of heuristics for biological sequence comparison," *BMC Bioinformatics*, **6** :31 (2005)]. Vous pouvez télécharger exonerate à http://www.ebi.ac.uk/~guy/exonerate/. Étudiez

[1] SGD=*Saccharomyces Genome Database*

la documentation du logiciel. Aligner avec `--model protein2genome` pour obtenir les coordonnées (rélatives) de l'intron dans la sortie (`--showtargetgff yes` recommandé).

4. Extraire la séquence de l'intron à partir des coordonnées de l'alignement et la séquence ADN de la région.

Mettez les 5 séquences introniques dans un seul fichier, avec noms Scer, Spar, ...

e. Pistes d'annotation pour l'intron (5 points). Utilisez BigFoot sur les séquences introniques, et créez des pistes WIG à partir des sorties `mpd` et `pred` comme avant. Faites attention à l'orientation des brins.

1.3 Exercices à option ($\leq$ 20 points boni)

a. Automatisation (12 points boni). Implantez un ensemble d'outils et performez les étapes §1.2**d**–**e** pour tous les introns du génome. Il y a 281 gènes avec introns ($\leq$ 300 introns au total). Il faut écrire les scripts pour compiler les séquences d'ADN des introns pour un identifieur SGD donné, ainsi que générer l'annotation pour les introns (lancer BigFoot à la ligne de commande + produire les pistes WIG).

b. Score de parcimonie (12 points boni). Écrire un outil qui produit une piste d'annotation WIG avec le score de parcimonie de chaque nucléotide alignée (reconstruction ancestrale à partir de la sortie MPD de BigFoot).

c. Outil de segmentation (12 points boni). Concevoir et implanter un outil de segmentation à partir des probabilités dans la sortie de BigFoot. Le programme devrait écrire une piste dans le format BED.

d. Adapter un outil (20 points boni). Adaptez un outil existant pour l'annotation de conservation ans les 5 espèces à partir des alignments MPD de BigFoot : SiPhy [http://www.broadinstitute.org/genome_bio/siphy], SCONE [http://genetics.bwh.harvard.edu/scone/], ou PhastCons [Hubisz, Pollard, Siepel "PHAST and RPHAST : phylogenetic analysis with space/time models," *Briefings in Bioinformatics*, **12** :41–51 (2010)].

1.4 Logistique

Ce devoir est destiné à des équipes de deux, mais vous avez le droit de travailler seul. Le devoir vaut 20 points et vous pouvez avoir 20 points de boni pour des exercices de programmation de §1.3.

Remise. Remettez les fichiers suivants. (1) Fichiers de Bigfoot : alignements ou séquences à l'entrée, sortie `.pred`, `.mpd`, `.log` ; (2) pistes d'annotation (`.wig`) ; (3) code implantée ; (4) rapport (format libre, en texte ou PDF) incluant un journal de ce que vous avez fait exactement : URLs de fichiers téléchargés, noms de fichiers locaux, paramètres à la ligne de commande, options choisies dans les logiciels. Mettez tous les fichiers dans un seul archive tar (p.e., `tar cf tp1.tar devoir1/; gzip tp1.tar`), et soumettez comme pièce jointe dans un e-mail à csuros@iro.umontreal. ...

Date limite. Soumettez avant 21 :00 le 27 février.