

**INF 4500 Bioinformatique** Professeur: Anne Bergeron.  
Examen intra à remettre le 21 octobre 2008.

**Directives.** Ceci est un examen individuel. Vous avez le droit de consulter toutes les ressources disponibles, y compris la littérature scientifique et vos collègues étudiants. Votre rapport doit toutefois citer vos sources et clairement démontrer que vous en êtes le seul auteur.

**Question 1** 50 points.

a) Choisissez deux protéines homologues de l'humain et de la souris qui soient très similaires. Ces protéines ne doivent pas être des protéines codées dans les mitochondries, ou discutées dans le cours ou dans les ateliers. Expliquez comment vous avez procédé (recherche sur le Web, discussion avec un biologiste, utilisation d'outils de recherche spécialisés, etc.). Décrivez brièvement son rôle dans l'organisme, en donnant les références pertinentes. Donnez leur numéro d'accession Genbank, et leur taille en acide aminés.

b) Pour chacune des deux protéines, donner la localisation, dans le génome, du gène (locus) qui code pour la protéine. C'est-à-dire les numéros des chromosomes, les adresses de départ et de fin des locus, et la longueur totale des locus. Quelle proportion de ces locus code pour vos protéines? Détaillez vos calculs.

c) Donnez un alignement optimal des deux protéines, en précisant le logiciel que vous avez utilisé et les paramètres correspondants, dont le nom de la matrice de score. Quel est leur **score** de similarité (il ne s'agit ici ni du pourcentage d'identité, ni du pourcentage de similarité, comme ceux donnés par le logiciel Align, par exemple).

d) Trouvez une protéine homologue aux deux précédentes qui n'appartient pas à un mammifère (ie. poisson, insecte, oiseau, champignon, levure, etc). Expliquez comment vous avez procédé. Donnez son numéro d'accession Genbank, et sa taille en acide aminés. Dites de quel organisme elle provient, et donnez son score de similarité et son pourcentage d'identité, basé sur un alignement optimal, avec chacune des deux précédentes.

e) Trouvez une protéine homologue aux deux premières qui appartient à un primate (chimpanzé, gorille, bonobo, par exemple). Donnez son numéro d'accèsion Genbank, et sa taille en acide aminés. Expliquez comment vous avez procédé.

f) Donnez un alignement multiple des quatre protéines trouvées en a), d) et e), en couleurs si possible. Commentez brièvement la qualité de cet alignement.

**Question 2** 50 points.

Écrivez le pseudo-code d'un algorithme qui répond aux spécifications suivantes. Votre texte devrait être assez précis pour que votre technicien – formé au Cegep de la Belle Province – puisse être en mesure d'implanter le logiciel sans difficultés.

En entrée, le logiciel reçoit une suite  $S$  de 32 nucléotides et un chromosome  $C = c_1 \dots c_K$  en format FASTA. Vous devez décider si, oui ou non, il existe une occurrence approximative **unique** de  $S$  dans le chromosome  $C$  qui soit à une distance d'édition inférieure ou égale à 2 de  $S$ . La taille  $K$  du chromosome  $C$  peut atteindre plusieurs centaines de millions de nucléotides.

La réponse est 'non' dans deux cas. Soit aucun facteur de  $C$  n'est à une distance inférieure ou égale à 2 de  $S$ , soit il existe deux facteurs non-chevauchants qui sont à une distance inférieure ou égale à 2 de  $S$ . Deux facteurs  $f = c_i \dots c_j$  et  $g = c_n \dots c_m$  sont non-chevauchants si  $j < n$  ou  $m < i$ .

Par exemple, si  $S = aca$  et  $C = gtcacacatttgct$ , la réponse est 'oui', mais si  $S = aca$  et  $C = gtcacacatttgacat$ , la réponse est 'non'.

**Bonus** Un algorithme correct et simple est suffisant pour avoir une bonne évaluation. Vous êtes toutefois invités à proposer des moyens pour accélérer significativement la recherche. L'inclusion de références pertinentes à la littérature scientifique est jugée très favorablement.