Projet CAML 2010 Sequençage de l'ADN

1) Etant donné une chaîne de nucléotides (une liste sur l'alphabet A T C G), construire les fonctions donnant leur traduction en suite d'acides aminés (triplets de nucléotides) et en suite de gènes en prenant des exemples artificiels (chromosomes de longueurs réduites et une dizaine de gènes définis arbitrairement par nn = 10 à 15 acides aminés).

2) On définit les opérations et ensembles suivants :
Complémentaire, c'est l'opération, pour un acide aminé, consistant à lire dans l'ordre inverse l'acide aminé correspondant (A avec T et C avec G) sur l'autre brin d'ADN. Ainsi co(ACT) = AGT car ACT est associé à TGA.
Ainsi les séparateurs reconnus par TGA, TAA, TAG, sont-ils associés à TCA, TTA, CTA.
Permuté, c'est la permutation circulaire, ainsi pr(ACT) = CTA

Les 64 trinucléotides peuvent se ranger en trois ensembles X_0 = {AAC, ...} comprenant AAA, TTT et les 20 autres tels que X_0 soit stable par complémentation, X_1 = {ACA, ...} formé par les permutés circulaires de X_0 excepté AAA et TTT, mais comprenant CCC. X_2 = {CAA, ...} est formé des 21 permuté circulaires de X_1 excepté CCC, mais avec GGG.
Construire des fonctions produisant ces 3 ensembles et vérifier que X_1 et X_2 sont complémentaires l'un de l'autre.
L'ordre de grandeur des probabilités d'être dans X_0 est très environ 0.5 dans la chaîne chromosomique, celle de X_1, 0.3, et 0.25 pour X_2.
En réalisant des mutations (modification aléatoire d'un nucléotide = d'une lettre) à partir d'une chaîne uniquement composée de codons de X_0, on arrive expérimentalement après une "quelques" mutations à 3 probabilités voisines (en moyenne 200).
On souhaite le vérifier en créant des chromosomes aléatoires (leur longueur peut se limiter à une vingtaine) et les mutations peuvent porter sur un trinucléotide seulement, on testera après avec le cross-over entre deux chromosomes à deux sites quelconques en coupant éventuellement au sein des trinucléotides et avec la mutation ne portant que sur un nucléotide.
Entre deux chromosomes, on appelle croisement ou crossover, leur remplacement par les deux chromosomes obtenus en échangeant un segment entre deux positions aléatoires.
Donner le nombre moyen (obtenu au cours d'une centaine d'expériences) d'opérateurs génétiques nécessaires, suivant leurs types, pour obtenir 3 fréquences voisines à 0.05 près (est-ce plus rapide en utilisant ou non des cross-over ?).
3) On considère une population homogène pour un certain locus (lieu) où le gène P est présent avec la fréquence p0. Soient μ et ν les probabilités respectives qu'en ce locus P mute en M et M en P.
Donner la relation de récurrence entre les probabilités p_n et p_{n-1} de présence du gène P aux générations n et n-1.
Donner une expression directe de p_n et, à l'aide d'une fonction, étudier la vitesse de sa convergence.
Etudier la dans le cas de la diploïdie, si P est le gène du père et M celui de la mère, sans qu'il y ait dominance mais avec $\mu = \nu$ petit mais non nul. n=20 suffit souvent pour une bonne limite $\nu/(\mu+\nu)$

Réponse $p_n = (1-\mu)p_{n-1} + \nu(1-p_{n-1}) = (1-\mu-\nu)^n p_0 + \nu(1-(1-\mu-\nu)^n)/(\mu+\nu)$

4) La probabilité de mutation est environ 10^{-5} par gène, l'homme avec 25000 gènes est donc porteur en moyenne d'une mutation une fois sur 4.
L'actuelle thèse "neutraliste" corrige le darwinisme en étudiant le fait que sur des populations réduites, les mutations ont plus de chances de se fixer de façon régulière. Afin d'expérimenter on considère une population de np individus identiques déterminés chacun par un génome de m lettres A, B, C, D ... ou m triplets symbolisant des gènes. On simule une évolution accélérée polyploïde assez semblable à l'autofécondation, en passant d'une génération à une autre de la manière suivante :
Chaque individu va produire un seul fils obtenu en remplaçant un segment par le segment correspondant d'un partenaire pris au hasard, puis on fait subir à cette chaîne une mutation aléatoire. Il n'y a donc pas de probabilité à fixer. La population des np fils va constituer la génération suivante.
On étudie les n premières mutations, chacune étant déterminée par une lettre et une position, leur fixation sera mesurée par la moyenne des fréquences de ces n mutations au cours des générations suivantes. Construire toutes les fonctions nécessaires à l'étude de np (effectif de la population), ng (nombre de générations) $\rightarrow$ probabilité f de fixation d'une mutation. Pour ng fixé par exemple à 100, étudier la convergence de n $\rightarrow$ f.