Atelier : Rapprocher les gènes
Robert Bilinski
Camp AMQ 2005 – niveau Cégep
Université du Québec à Montréal

1. Introduction :

Pour diverses raisons, il est important de pouvoir mesurer la distance entre 2 gènes. Par exemple, pour étudier la provenance d'un échantillon d'influenza dans le but de savoir quel vaccin injecter ou produire, il faut comparer l'ADN de l'échantillon à l'ADN contenu dans les banques de données et utiliser le vaccin déjà développé. En génie génétique, pour contrôler la qualité des gènes produits, il faut être capable de comparer l'ADN réellement produit avec les séquences d'ADN planifiées. En génétique des populations, on veut connaître la similarité entre 2 espèces (tribus d'hommes, meute de loup, souche de bactérie…). On peut espérer créer un arbre généalogique de brins d'ADN et des espèces d'où ils proviennent.

Les gènes sont des regroupements d'acides désoxyribonucléiques (ADN). Il y en a 4 types naturels : L'Adénine, la Thymine, La Guanine et la Cytosine. On utilise les lettres A, T, G, et C pour abréger. Les ADN se regroupent et forment des chaînes en forme de double hélice. La séquence sur une des hélices est le négatif de la séquence de celle que l'on retrouve sur l'autre. En effet, quand il y a un C sur l'une, à l'endroit correspondant sur l'autre, il y aura un G et vice-versa. Il en est de même pour les A et les T. Ainsi, on représente les brins d'ADN par la séquence de A, T, C et G que l'on retrouve sur un des 2 brins. Pour illustrer la chose, on aura

Hélice 1 : …ATTTCGGGACCGAAAT…
Hélice 2 : …TAAAGCCCTGGATTTA…

Mais on ne gardera que …ATTTCGGGACCGAAAT… dans la version écrite. Dans ce contexte, on ne pense plus aux lettres comme des acides (donc des molécules), mais comme des lettres dans un alphabet. On parle alors de code génétique quand on parle de tout l'ADN dans un organisme. A, T, C et G sont appelés les bases du code génétique.

L'organisme utilise les séquences de ces ADN pour stocker de l'information (couleur des yeux, forme du nez, taille, sexe, création des hormones, de protéines, …).

À travers le temps, ces séquences changent. On parle alors de mutation. Il y a plusieurs types de mutation. Certaines plus compliquées (translocation, amplification, etc), et donc plus rares, peuvent être considérées comme une combinaison de 3 mutations de bases : la délétion, l'insertion et la substitution. Pour pouvoir modéliser mathématiquement ces mutations, il faut introduire aux 4 lettres A, T, C et G de base, le symbole _

qui représente un espace vide, ou mathématiquement la base vide. Dans ce domaine, on utilise les lettres X et Y pour identifier des bases inconnues. Les lettres minuscules, a et b par exemple, servent à nommer des séquences. En ajoutant des indices au nom d'une séquence, on peut identifier une base spécifique dans une séquence.

Pour décrire les mutations, il existe plusieurs standards. Nous utiliserons ici le suivant. Un couple (X , Y) indique que l'on a changé la lettre X par la lettre Y. Ainsi :

– (_ , X) désigne l'insertion du nucléotide X qui n'était pas là avant

– (X , _) désigne la délétion du nucléotide X qui n'est plus là après

– (X , Y) désigne la substitution du nucléotide X par le nucléotide Y.

Par contre, on n'indique pas que (X , Y) par exemple a eu lieu à la 5^{e} base de la séquence a, car quand on essai d'identifier une mutation, la représentation d'une séquence n'est plus unique. En effet, pour savoir quelle mutation a eu lieu, il faut envisager toutes les possibilités suivantes et bien d'autres :

a = ATCGGG
= A_TCGGG
= AT_CG_ _G_G

Questions de réflexion :

#1) Existe-t-il toujours une suite de mutations entre 2 gènes?

#2) Combien de « progénitures possibles » a une séquence d'ADN donnée?

#3) Peut-on affirmer avec certitude qu'une séquence d'ADN provient d'une autre?

Conclusion 1 :

2. Aligner des séquences d'ADN

On appelle « Aligner 2 séquences » essayer de trouver une manière de faire correspondre les bases de la première avec les bases de la seconde.

Par exemple, pour aligner a = AT et b = AAGT. On a plusieurs alignements possibles :

A A G T , A A G T , A A G T … … et A A G T .
A T _ _ A _ _ T _ _ A T _ A _ T

Certains sont visibles « meilleurs » que d'autres.

Il existe plusieurs approche pour arriver à aligner des séquences :

Un alignement « local » essai de trouver le plus de couples correspondants identiques entre les 2 séquences comparées.

Un alignement « global » essai de trouver un équilibre entre les couples identiques (en avoir le plus possible) et les couples différents (en avoir le moins possible)

Questions de réflexion :

#4) Que veut-on dire par meilleur alignement?

#5) Comment faire un alignement global (plus simple)?

#6) Comment faire un alignement local?

#7) Est-ce que toutes les mutations sont équiprobables?

Exercice :

#1) Aligner CCTG et ATCTTGGA