

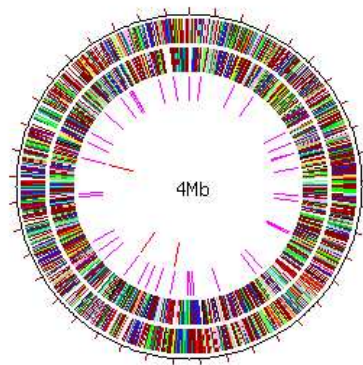
## Exercices de programmation dynamique

### ACCGGTTNTTCA

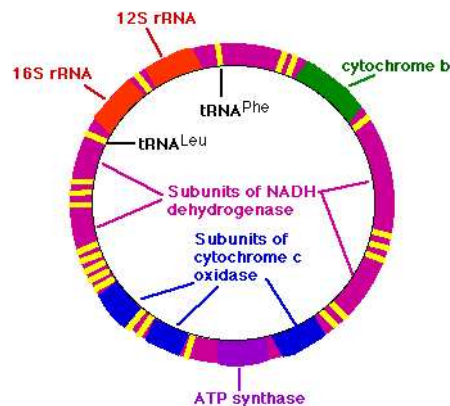
Lors de l'obtention d'une séquence nucléique ou protéique, il arrive que certaines positions ne soient pas déterminées avec certitude, à cause notamment des erreurs ou de l'ambiguïté des techniques de séquençage. Dans ce cas, la position est notée N pour les acides nucléiques, et X pour les séquences protéiques.

Proposez un algorithme d'alignement permettant de gérer les séquences avec des positions indéterminées, telles que N ou X.

### Comparaison de génomes circulaires



génomme de *Bordetella pertussis*  
(4 millions bp)



génomme mitochondrial humain  
(16 569 bp)

Les algorithmes vus en cours pour l'alignement s'appliquent à des séquences linéaires, avec un début et une fin, comme un gène ou un génome linéaire. Mais il existe également des génomes circulaires. C'est le cas des génomes bactériens, et chez les eucaryotes des mitochondries et des chloroplastes.

Comment peut-on prendre en compte les génomes circulaires en adaptant les algorithmes d'alignement global (Needleman & Wunsch) et d'alignement local (Smith & Waterman) ?

### Recherche de répétitions

Étant donnée une séquence, une répétition est une sous-séquence qui apparaît au moins deux fois, de manière exacte ou approchée, sans se chevaucher. On souhaite écrire un algorithme qui détermine quelle est la répétition de meilleur score dans une séquence.

**Question 1.** Lequel des deux algorithmes d'alignement, global ou local, est-il le plus utile dans ce contexte ?

**Question 2.** Comment peut-on en modifiant le remplissage de la table de programmation dynamique résoudre le problème ?

**Question 3.** Peut-on facilement accélérer le remplissage de la table dans le cas de la recherche de répétitions ? Comment ?

### Réplication hasardeuse

Une bactérie est atteinte par un virus qui affecte la machinerie de la réplication de la sorte

- chaque A peut être remplacé par 3 A,
- chaque C peut être remplacé par 4 C,
- chaque G peut être remplacé par 4 G,
- chaque T peut être remplacé par 3 T.

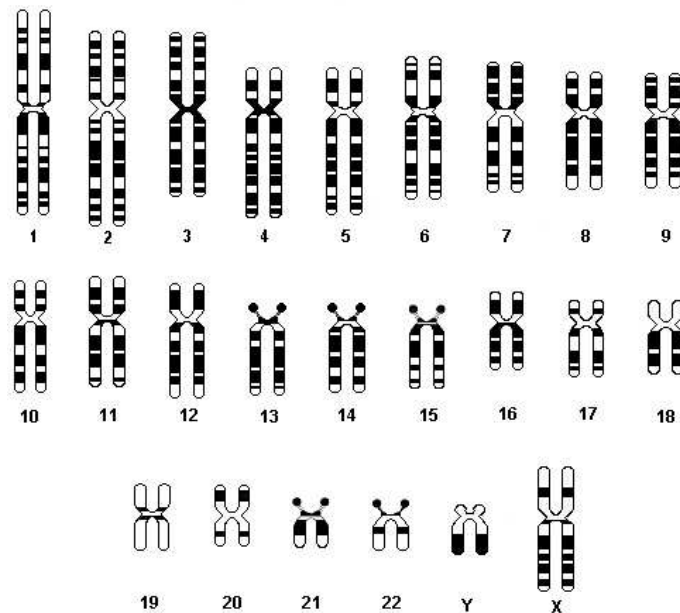
Notez que la multiplication des bases n'est pas systématique, et il se peut qu'à une position donnée la copie soit correcte.

**Question 1.** Donnez un algorithme efficace qui pour deux séquences  $u$  et  $v$  détermine si  $u$  peut être une version infectée de  $v$ .

Le virus a muté, et en plus d'induire une copie multiple d'une position, il est également possible que la base soit oubliée, provoquant une délétion.

**Question 2.** Modifiez l'algorithme précédent pour prendre en compte ce nouveau phénomène.

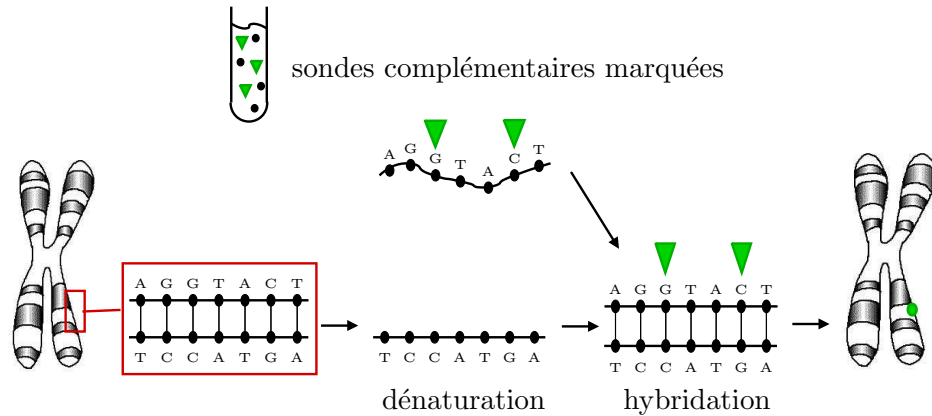
### Localisation des bandes des chromosomes



L'observation au microscope de chromosomes fait apparaître des bandes transversales plus ou moins foncées. La cause exacte de ce phénomène est mal comprise. Cela peut provenir du pourcentage en GC. Le génome humain, par exemple, avec ses 23 paires de chromosomes, compte un total de 862 stries. Les bandes sont répertoriées dans une nomenclature internationale. Chaque bras chromosomique est divisé, selon sa taille, en une à quatre régions; chaque région en bandes numérotées du centromère au télomère. Par exemple, la dénomination 6p12 désigne la deuxième bande de la région 1 des bras courts du chromosome 6.

Avant le séquençage, ces bandes ont servi à établir des cartes physiques du génome. Il serait maintenant intéressant d'établir une correspondance entre la topographie des stries et la séquence génomique, c'est-à-dire de déterminer les positions des bornes de chaque bande. Il n'existe pour cela pas d'approche systématique. La largeur des bandes donne seulement une indication très approximative sur la longueur de la séquence impliquée.

Les expériences de FISH (*Fluorescent in situ Hybridization*) peuvent néanmoins aider à résoudre ce problème. Un résultat de FISH permet de déterminer pour une position donnée du génome l'indice de la bande correspondante. Mais ces expériences sont sujettes à erreurs, et il faut donc un grand nombre d'observations pour en déduire des informations fiables.



Donnez un algorithme qui à partir d'un ensemble de données de FISH détermine lesquelles sont les plus fiables. On considère que l'ensemble le plus fiable est le plus grand ensemble ne contenant que des observations cohérentes entre elles.