

5. Les graphes de de Bruijn pour l'assemblage des génomes

Les séquenceurs d'ADN produisent de nombreuses petites séquences extraites de la longue séquence de quatre lettres A, G, C, T à laquelle on s'intéresse (un gène, un chromosome, etc.). Les petites séquences ont des parties communes qui déterminent leur assemblage correct. Une petite séquence

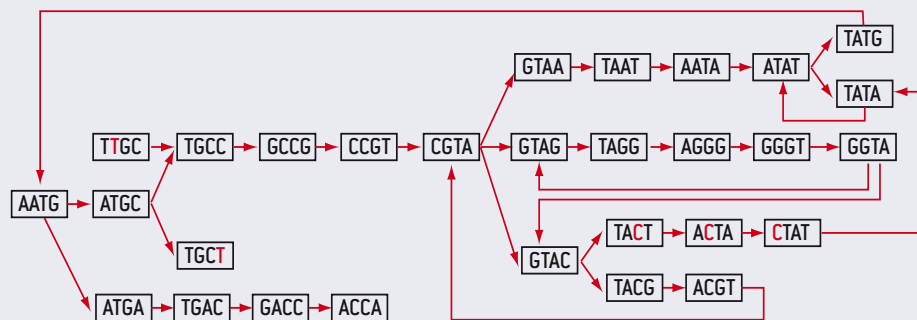
peut parfois s'assembler avec plusieurs, et on est donc face à un énorme puzzle : toutes les petites séquences produites par les machines doivent être utilisées, afin de ne déterminer qu'une seule grande séquence.

Depuis les travaux de Pavel Pevzner, Haixu Tang et Michael Waterman, les méthodes utili-

sées pour traiter ces énormes problèmes combinatoires sont fondées sur la recherche de chemins eulériens dans des graphes analogues à ceux de l'encadré 3. Les sommets de ces graphes (dont un exemple est représenté ci-dessous) sont les séquences données par la machine et les arcs indiquent les

paires de séquences ayant une partie commune (car l'une commence par ce qui est la fin de l'autre). Mener l'assemblage se ramène à trouver un chemin eulérien dans le graphe.

Le problème est rendu encore plus difficile par certaines erreurs présentes dans les résultats fournis par les machines à séquencer. Les lettres en rouge indiquent les endroits où il y a des erreurs de séquençage. La mise au point de ces algorithmes d'assemblages à base de graphes de de Bruijn fut une étape importante dans l'histoire du séquençage et aujourd'hui encore de nombreux travaux sont consacrés au perfectionnement de ces méthodes.



quadruplets possibles de 0 et de 1. En effet, le circuit passe deux fois par tout sommet (un triplet) ; une première fois, il le quitte en choisissant un 0, la seconde fois en choisissant un 1. Chaque quadruplet est obtenu une seule fois, sinon le circuit passerait deux fois par le même arc.

Donnons un exemple. Soit le circuit passant par les sommets 000–000–001–011–110–100–001–010–101–011–111–110–101–010–100. Sa notation compacte est la suite de de Bruijn d'ordre 4 notée 0000110010111101.

On vérifie qu'en prenant quatre symboles consécutifs, on obtient les 16 suites possibles de 4 symboles 0 ou 1 dans l'ordre suivant : 0000–0001–0011–0110–1100–1001–0010–0101–1011–0111–1111–1110–1101–1010–0100–1000.

La démonstration complète que, par cette méthode, on trouvera des suites de de Bruijn d'ordre n pour tout n est presque terminée. Il suffit d'évoquer un célèbre résultat de théorie des graphes dû à Leonhard Euler : **Théorème d'Euler** : Un graphe orienté possède un circuit empruntant tous les arcs une fois exactement (dénomé circuit eulérien) si et seulement si :

- il est connexe (on peut toujours joindre deux sommets en empruntant les arcs du graphe) ;
- à chaque sommet, le nombre d'arcs entrants est le même que le nombre d'arcs sortants.

Circuits eulériens

Pour les graphes considérés ici, la seconde condition du théorème est vérifiée, car de chaque sommet partent deux arcs, et en arrivent deux. Pour la première condition, il faut réfléchir un peu : partant de abc (cas $n = 3$), peut-on arriver à efg ? Oui, il suffit de suivre les arcs abc , bce , cef , efg . Pour n quelconque, l'idée s'adapte.

Les graphes de de Bruijn vérifient les conditions du théorème d'Euler. Ils ont donc tous des circuits eulériens, et donc pour tout entier n , il existe des suites de de Bruijn. Conclusion : quel que soit n , on peut faire le tour de cartes avec n spectateurs et un jeu de 2^n cartes. La méthode des circuits eulériens permet de trouver les suites de de Bruijn d'ordre n , pour tout n , et elle les donne toutes. Cela permet de les dénombrer. Il y en a exactement $2^{2^n - 1 - n}$.

Une multitude de résultats ont été démontrés concernant les suites et les

graphes de Nicolaas de Bruijn et leurs variantes. Ces résultats sont utiles par exemple en robotique (voir l'encadré 4) ou dans le domaine du traitement des séquences génétiques pour assembler les courtes séquences que les machines déchiffrent en grand nombre et qu'il faut recoller pour connaître, par exemple, la totalité d'un chromosome (voir l'encadré 5).

Le merveilleux livre *Magical Mathematics* de Persi Diaconis et Ron Graham (voir la bibliographie) consacre trois chapitres aux tours de magie fondés sur les suites de Nicolaas de Bruijn et décrit plusieurs autres applications de ces suites. On y trouve aussi des informations sur l'histoire du tour présenté au début d'article. Elle est intéressante, car elle montre qu'avant que les mathématiciens ne s'en mêlent, l'idée du tour était déjà comprise, mais n'était qu'imparfaitement mise en œuvre. La première version du tour fut exposée en 1919 par Charles Jordan, un fermier californien, éleveur de poulets, qui inventait et vendait des tours de magie. Cette première version n'était pas la meilleure possible : le magicien, qui utilisait pourtant un jeu de 32 cartes, interrogeait six spectateurs au lieu de cinq. Le rangement du jeu ne