

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

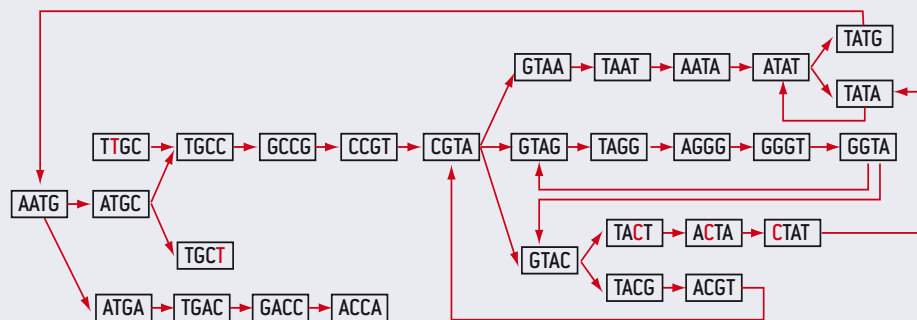
peut parfois s'assembler avec plusieurs, et on est donc face à un énorme puzzle : toutes les petites séquences produites par les machines doivent être utilisées, afin de ne déterminer qu'une seule grande séquence.

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 un 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

devait donc pas suivre une suite de de Bruijn. En 1930, une version intéressante du tour est décrite par deux autres inventeurs de tours, William Larson et T. Page Wright. Cette fois, ils manipulent un jeu de 52 cartes et demandent à trois spectateurs ayant retiré trois cartes consécutives si leur carte est un cœur, un carreau, un pique ou un trèfle. Disposant d'une information de 6 bits (2 bits par réponse) le magicien retrouve la coupe faite dans le jeu de 52 cartes (qui demande 5,7 bits d'information) et aurait même pu utiliser un paquet de 64 cartes (qui aurait nécessité 6 bits d'information).

Pour mener l'analyse mathématique de cette variante du tour, il faut utiliser des suites de symboles pris dans un alphabet à 4 symboles. C'est effectivement ce que la théorie s'est attachée à traiter. Elle considère la situation générale d'un alphabet de k symboles et des séquences de longueur n . Le problème est alors de savoir s'il existe une suite de symboles (de longueur k^n) qui donne les k^n suites de n symboles possibles quand on considère n éléments consécutifs de la suite. La réponse à ce problème généralisé est à nouveau *oui*. Nicolaas de Bruijn a même su dénombrer les solutions : pour tout entier $n \geq 1$ et tout entier $k \geq 1$, il existe des suites de de Bruijn de longueur k^n , et leur nombre est exactement : $(k!)^{k^n-1}/k^n$.

Aide au cambrioleur

Comme application amusante des suites de de Bruijn, il y a l'ouverture d'une serrure à 4 chiffres décimaux ne comportant pas de touche « entrée » : quand les quatre derniers nombres tapés sont bons, la serrure s'ouvre. Il semble qu'une attaque par la force brute exige d'essayer 10 000 combinaisons. Si on les entre les unes après les autres par ordre croissant (0000, 0001, 0002, ..., 9999), il faut taper 40 000 fois sur le clavier. En revanche, si on utilise la bonne suite de de Bruijn, il suffira de taper 10 003 touches : les 10 000 de la suite elle-même, suivies des trois premiers chiffres de la séquence pour parcourir le cycle complet.

Parmi les nombreux résultats que les mathématiciens continuent de rechercher

et de démontrer concernant les suites de de Bruijn, citons-en deux récents.

En 2011, Veronica Becher et Pablo Ariel Heiber se sont demandé si, disposant d'une suite de de Bruijn d'ordre n sur un alphabet de k symboles, il était possible de la prolonger (en en multipliant la longueur par k) de façon à obtenir une suite de de Bruijn d'ordre $n+1$. Ils ont établi que (a) si $k \geq 3$, la réponse est toujours *oui* ; (b) si $k = 2$, alors on ne peut jamais prolonger une suite de de Bruijn d'ordre n en une suite d'ordre $n+1$, mais on peut toujours la prolonger en une suite d'ordre $n+2$.

Plutôt que d'avoir une suite de 0 et de 1 qui, lorsqu'on fait glisser sur elle une fenêtre de largeur n , énumère toutes les séquences de longueur n possibles (c'est la définition d'une suite de de Bruijn d'ordre n , pour l'alphabet à deux éléments 0 et 1), on peut chercher à énumérer par la même fenêtre glissante uniquement les séquences de 0 et de 1 pas trop denses, par exemple uniquement celles contenant au plus m fois le 1. C'est la question que J. Sawada, B. Stevens, A. Williams se sont posée en 2011. Leur nombre est : $C_n^0 + C_n^1 + \dots + C_n^m$.

Il s'agit de la somme des $m+1$ premiers éléments de la ligne $n+1$ du triangle de Pascal. Rappelons que le coefficient $m+1$ de la ligne $n+1$ du triangle de Pascal (aussi appelé coefficient du binôme) vaut $C_n^m = n!/(m!(n-m)!)$. Peut-on, comme dans le cas de l'énumération complète de toutes les séquences de longueur n , énumérer de la façon la plus économique possible les séquences ayant au plus m '1', c'est-à-dire en déplaçant une fenêtre de largeur n , sur une suite dont la longueur est le nombre écrit plus haut ?

La réponse, pas évidente du tout, est : oui, il existe des suites de de Bruijn pour l'énumération des séquences de longueur n et de densité au plus m .

Des versions en deux dimensions de ces problèmes d'énumération économiques ont aussi été étudiées (voir l'encadré 4) et servent en robotique. De la magie de music-hall aux difficiles problèmes combinatoires, le chemin n'est pas long, et les résultats mathématiques que l'on découvre en chemin concernent des domaines inattendus et variés, avec des applications concrètes.

L'AUTEUR



J.-P. DELAHAYE est professeur émérite à l'Université de Lille et chercheur

au Laboratoire d'informatique fondamentale de Lille (LIFL).

BIBLIOGRAPHIE

M. Dow, *Minimal arrays containing all sub-array combinations of symbols : De Bruijn sequences and tori*, 2014 : <http://tiny.cc/9y96fx>

J. W. Klop, Nicolaas Govert de Bruijn (1918-2012) - Mathematician, computer scientist, logician, *Indagationes Mathematicae*, vol. 24, pp. 648-656, 2013.

P. Diaconis et R. Graham, *Magical Mathematics*, Princeton University Press, 2012.

J. Sawada et al., *De Bruijn sequences for the binary strings with maximum density*, WALCOM'11 Proceedings of the 5th international conference on WALCOM: algorithms and computation, pp. 182-190, Springer, 2011.

V. Becher et P. A. Heiber, *On extending de Bruijn sequences*, *Information Processing Letters*, vol. 111, pp. 930-932, 2011.

C. Flye Sainte-Marie, *Solution à la question n° 48*, *L'Intermédiaire des Mathématiciens*, vol. 1, pp. 107-110, 1894.



Retrouvez la rubrique Logique & calcul sur www.pourlascience.fr