Génomique et Optimisation Combinatoire

Alain Guénoche
Institut de Mathématiques de Luminy
163 Av. de Luminy
13288 Marseille cedex 09
email: guenoche@iml.univ-mrs.fr

- Conférence invitée -

1 Introduction

La Biologie Moléculaire, est l'étude des mécanismes du vivant à travers l'étude et la comparaison des molécules présentes ou passées dans les différents organismes qui peuplent la planète. Nous nous intéresserons à quelques problèmes informatiques qui découlent de l'analyse des génomes ; on parle de génomique et même de post-génomique, quand on s'intéresse au décryptage des séquences biologiques, aux fonctions des protéines et à leurs intéractions.

Les macromolécules de toutes les cellules de toutes les espèces sont des enchaînements d'un petit nombre de motifs, si bien que l'on peut les décrire par des séquences ou chaînes de caractères sur des alphabets finis.

- Il y a des séquences d'ADN. Cet acide est présent dans chaque cellule sous la forme de deux suites de motifs biochimiques, les desoxiribo-nucléotides, également appelés bases. Ces motifs sont au nombre de 4, si bien qu'on leur fait correspondre un alphabet à 4 lettres A,C,G,T. Chaque suite, applée brin, est orientée et les deux brins sont de sens opposés et complémentaires; à un A (resp. G) dans l'un correspond un T (resp C) dans l'autre ; ce sont les paires de Watson-Crick. Les deux brins sont liés (physiquement) et forment une double hélice. Enfin un ADN peut être d'un seul tenant, parfois circulaire comme chez certaines bactéries, ou composé de morceaux, les chromosomes. Chaque brin contient des gènes. Les gènes sont des régions qui contiennent l'information nécessaire pour synthétiser les protéines ; on dit qu'un gène code pour une protéine. Chez l'homme il y a de l'ordre de 30 000 gènes, beaucoup moins qu'initiallement prévu.
- Il y a des séquences d'ARN. L'alphabet est encore à 4 caractères A,C,G,U correspondant aux ribo-nucléotides. Ces ARN ont plusieurs fonctions; les plus classiques,

sont les ARN messagers, obtenus par transcriptions (copie) des gènes. Après l'opération d'épissage (chez les eucaryotes), qui a pour fonction de concaténer les parties codantes, ils passent au travers des ribosomes qui interprêtent les ARN par groupe de 3 bases, les codons. A chaque codon correspond un acide aminé spécifique, désigné par le code génétique, qui est ajouté à la suite pour la synthèse des protéines ; c'est la traduction.

- Il y a des séquences protéiques. Les éléments de base sont des acides aminés, encore appelés résidus, au nombre de 20. Ils sont désignés par les lettres de l'alphabet, toutes sauf B,J,O,U,X,Z. Ces acides aminés partagent certaines propriétés (acide/basique, hydrophile/hydrophobe, etc) et occupent dans l'espace des volumes différents, si bien qu'il en est de plus semblables que d'autres. Ces proximités entre acides aminés peuvent être quantifiées de différentes manières. Les séquences protéiques vont d'une centaine d'acides aminés à quelques milliers, généralement quelques centaines.

L'essentiel de mon exposé consistera à présenter certains problèmes de biologie moléculaire qui peuvent être abordés par des méthodes combinatoires et plus particulièrement des méthodes couramment employées en Analyse des Données et en Recherche Opérationnelle, comme l'optimisation discrète, l'ordonnancement, la modélisation par des graphes et les algorithmes afférants.

2 Quelques problèmes

Le séquençage
Le premier problème à propos de ces séquences a été de les constituer ; c'est ce qu'on appelle le séquençage réalisé grâce à des collaborations internationales. De nos jours un tout petit nombre d'organismes ont leur ADN complètement séquencé (ou presque). Comme la méthode principalement utilisée [Sanger et al. 1977] ne permet d'obtenir que la séquence d'un fragment de longueur 400 à 500 bases, on butte sur un premier problème combinatoire, l'assemblage des différents fragments. Ceux-ci étant plus ou moins tirés au hasard au long du génome, et leurs chevauchements étant détectés par identité d'une éventuelle partie commune, on reconnaît là un problème de graphe d'intervalle : les sommets sont les fragments séquencés, les arcs les paires qui se chevauchent. S'il s'agit d'un graphe d'intervalles, les fragments forment un seul morceau contigu, correspondant à la séquence ; sinon, quels arcs faut-il ajouter ou retrancher pour obtenir un tel graphe ?

Une fois la séquence acquise, arrive l'analyse proprement dite, qui consiste à annoter ce génome, c'est à dire à reconnaître des zones fonctionnelles, comme les parties codantes, en désignant le début et la fin des gènes, ainsi que les sites où se fixent les activateurs et inhibiteurs de la transcription. Les meilleures méthodes connues relèvent principalement des modèles probabilistes, basés sur les variations de la fréquence des mots chevauchant de longueur fixée, en particulier par la recherche de chaînes de Markov cachées ou de l'algorithmique des chaînes de caractères (string matching problem). Bien que ces méthodes soient de type combinatoire, nous ne les considèreront pas ici. Par contre, une fois identifiées les séquences correspondant à des fonctions identiques chez différents organismes, on retrouve des problèmes auxquels s'appliquent l'arsenal bien connu de l'Optimisation Discrète et de la Programmation Mathématique. Nous continuons à présenter les problèmes, tels qu'ils se posent en Biologie :

L'alignement et la comparaison des séquences
Historiquement, la comparaison de deux séquences de nucléotides m et n, s'est faite à l'aide d'une distance d'édition. En définissant des opérations d'édition (substitution, insertion, suppression) munies d'une fonction additive de coût, la distance entre séquences est définie comme le plus petit coût d'une suite d'opérations qui permet de transformer m en n. Cette suite construit un positionnement de m par rapport à n. Sa détermination est faite par programmation dynamique ; elle a même été redécouverte à cette occasion. Pour des séquences protéiques, on utilise la notion duale de score, et un algorithme du même type, avec des matrices de coût de substitution particulières suivant la finalité de la comparaison.

Mais pour plus de deux séquences, les comparaisons par paires ne respectent pas la transitivité des positionnements. En particulier pour les problèmes d'évolution, pour lesquels la mise en évidence des mutations est essentielle, on procède à un alignement multiple des séquences. Une position verticale dans l'alignement, un site, présente toutes les variations actuelles d'un nucléotide ou d'un acide aminé ancestral. Dès lors on peut comparer les séquences en dénombrant des événements mutationnels.

Comparaison des ordre des gènes sur deux espèces
Depuis Darwin, il est admis que tout ensemble d'espèces vivantes dérive d'un ancêtre commun. En considérant un alignement multiple de séquences homologues - qui proviennent toutes d'une même séquence ancestrale - on ne peut en douter ; la conservation de longues zones très similaires ne peut être due au hasard. Mais pour les gènes communs à un ensemble d'espèces, l'ordre n'est généralement pas conservé. L'un des problèmes combinatoires que nous décrirons est celui de la comparaison de ces ordres, afin d'établir un autre type de distances mutationnelles.

On applique le principe de parcimonie qui veut que les différences observées soient le résultat du plus petit nombre de transformations qui, à partir d'un ordre unique, permettent d'obtenir différents ordres. Si la seule transformation considérée est la transposition de deux gènes consécutifs, le problème est le même que celui de la recherche d'un ordre médian à un profil d'ordres totaux ou d'un ordre total à distance minimum d'un tournoi. Mais si les transformations sont des déplacements ou des retournements d'intervalles de gènes consécutifs, le problème n'est plus si simple [3]. Plus précisemment, si l'on considère des permutations signées (le signe + ou - correspond au brin porteur de chaque gène) la distance mutationelle se calcule en temps linéaire, mais si l'on considère des permutations non signées, le calcul des distances est NP-difficile.

Classification fonctionnelle des protéines
C'est l'un des sujets les plus brulants de la génomique actuelle. En séquençant des génomes complets, on a identifié la quasi totalité des gènes de plusieurs organismes ; leur nombre varie grandement. Certains étaient connus et impliqués dans des réactions biochimiques en cascade, appelées voies métaboliques. Mais on s'est apperçu que plus de la moitié des gènes ne correspondait à rien de référencé dans les banques de séquences et qu'on ne savait rien de leur fonction. On a bien sûr cherché qu'elles étaient les protéines de fonction connue les plus similaires, mais ceci n'a pas donné grand chose, jusqu'à ce qu'on dispose de données différentes :

Des données d'interaction qui permettent de savoir si deux protéines interagissent directement l'une sur l'autre, ou si l'une interagit sur un complexe contenant l'autre ;

Des données d'expression mesurées à l'aide de puces à ADN. Elles permettent d'évaluer instantanément la quantité de tous les ARNs exprimés dans une cellule. L'exprérience peut se renouveler dans les mêmes conditions expérimentales à des instants successifs, dans des conditions différentes.

Quelque soient ces données, elles permettent de mesurer des distances entre protéines et l'information se résume à un graphe complet valué par ces distances. On peut alors appliquer des méthodes de partitionnement de graphes seuils pour construire soit des classes, soit des partitions à nombre de classes fixé qui optimisent certains critères. On retrouve des problèmes typiques d'optimisation combinatoire, comme Partition en cliques, Coloration optimale, etc [2]. Ces problèmes étant NP-difficiles et les données généralement très importantes, on développe des heuristiques et des meta-heuristiques qui permettent de construire des classes ou des partitions optimisées [1].

Références

[1] A. Guénoche, Partitions optimisées selon différents critères. *Mathématiques, Informatique et Sciences humaines*, 2002, à paraître.

[2] P. Hansen et B. Jaumard, Cluster analysis and mathematical programming, *Mathematical Programming* 79:191–215,1997.

[3] J. Kececioglu et D. Sankoff, Exact and approximation algorithms for the inversion distance between two chromosomes. *Proceedings 4th-Symposium Combinatorial Pattern Matching*, A. Apostolico et al. (Eds.), Springer-Verlag, 1993, 87-105.