Informatique et séquençage

5



L'activité de séquençage consiste à transformer de la matière en information. Au cours de ce processus, l'outil informatique intervient pour conserver et organiser les données recueillies, les analyser au moyen de grandes quantités de calcul et les communiquer de manière à faire connaître les résultats.

Conserver et organiser l'information

Dans le processus de séquençage interviennent de nombreuses étapes de fabrication portant sur une grande quantité d'échantillons. Chacun de ceux-ci doit rester identifiable tout au long du processus. Concrètement, cela signifie qu'il est nécessaire de pouvoir établir qu'une donnée a été déterminée à partir d'un fragment d'ADN parfaitement identifié, et qu'elle l'a été à l'issue d'une suite définie d'opérations expérimentales. On utilise pour cela des bases de données qui servent également à établir les tableaux de bord du Genoscope et à planifier ainsi ses activités. La quantité globale d'information générée par les expériences menées au Genoscope est de l'ordre de 10 milliards d'octets* (gigaoctets) par jour.

De grandes quantités de calculs

Grâce à des moyens de calculs importants qui mettent en œuvre des techniques d'analyse d'image et de compression de données, on réduit et l'on organise la grande quantité de données brutes générées quotidiennement. Cette capacité de calcul permet aussi de reconstituer la séquence de longues régions du génome à partir d'éléments beaucoup plus petits. On est ainsi conduit à reconstituer des puzzles dont chacune des pièces est une séquence que l'on comparera à toutes les autres. Pour un génome bactérien entier, par exemple, la reconstitution du puzzle peut demander plus de deux cent mille milliards de comparaisons de caractères.
Les séquences reconstituées doivent ensuite être comparées à celles qui ont été déterminées par des milliers de chercheurs à travers le monde, avant d'être stockées dans des bases de données internationales : ces comparaisons sont aujourd'hui le meilleur moyen que nous ayons pour attribuer une fonction biologique aux séquences que nous avons déterminées.

Communiquer

Les données produites sont mises à la disposition des autres membres de la communauté scientifique internationale. Réciproquement, le Genoscope réactualise chaque jour les bases des données produites ailleurs dans le monde, via Internet. Quotidiennement, le Genoscope met ainsi de l'ordre de plusieurs millions d'octets de données nouvelles à disposition sur le réseau, et recueille des millions d'octets représentant les nouvelles données établies par des biologistes du monde entier. À cette fin, nous exploitons une connexion au réseau Internet qui permet d'échanger 10 millions de bits (Mbits) – plus d'un million d'octets - par seconde.

En quelques chiffres

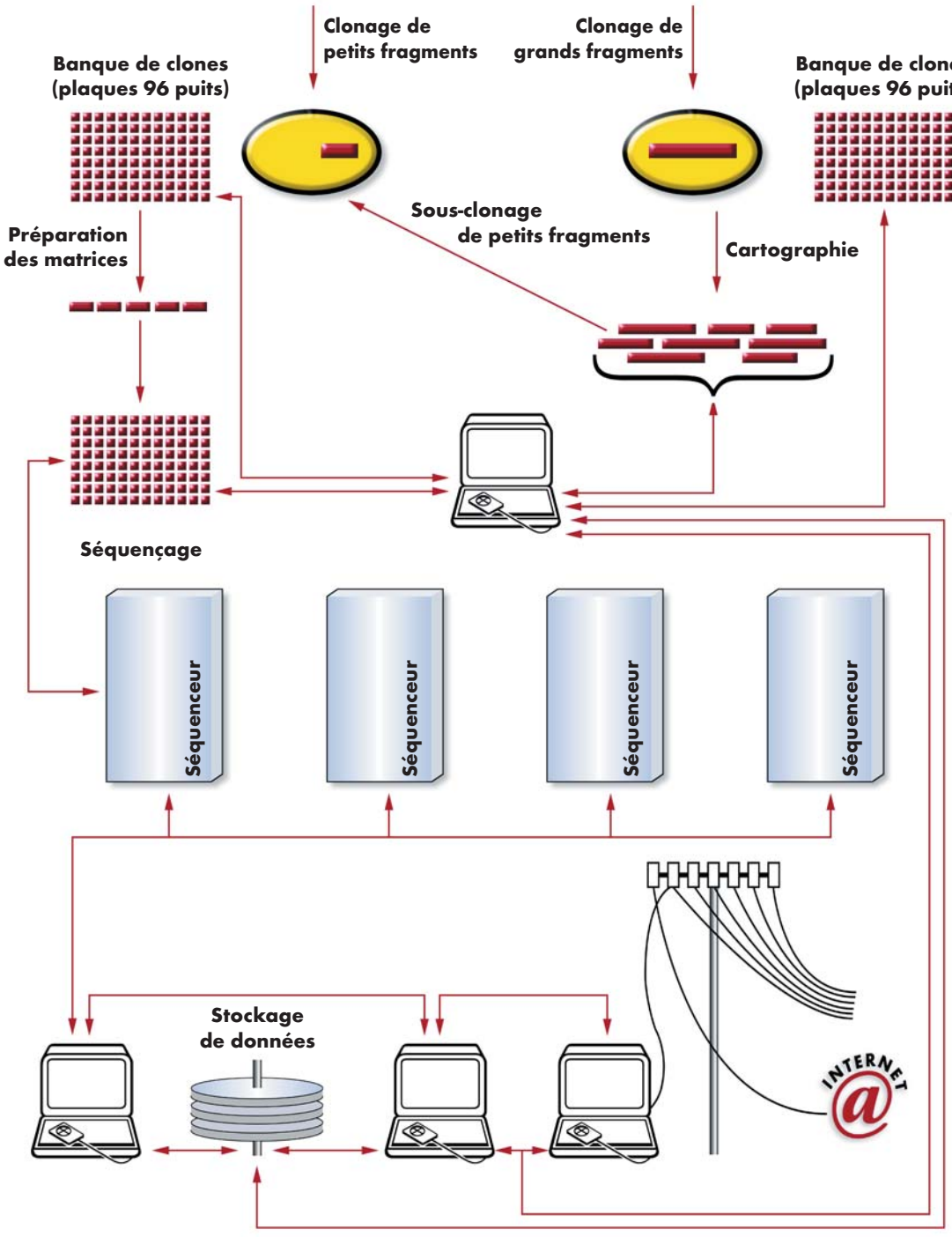
Le débit sur les lignes du réseau interne du Genoscope est de 1 à 100 millions de caractères par seconde suivant les équipements qu'elles desservent. La capacité totale de stockage sur disques est de 4 000 milliards d'octets ; un robot d'une capacité totale de stockage de 100 000 milliards d'octets assure les sauvegardes des données. La puissance de calcul a été déterminée de manière à ce qu'il soit possible de comparer chacune des séquences produites quotidiennement à l'ensemble des séquences connues à ce jour. *(suite au dos)*

** L'octet est l'unité de base de l'information numérique. Un octet, qui peut être comparé à un caractère, vaut huit bits (un bit c'est 0 ou 1). Un mégaoctet (Mo), c'est un million de caractères, un gigaoctet (Go), un milliard de caractères. Un livre de poche compte à peu près 375 000 caractères, donc un mégaoctet vaut trois livres de poche standard.*

Informatique et séquençage (suite)

5





Intégration de la gestion informatique des activités de cartographie et de séquençage du Genoscope