Article 22. Crise du stockage: conserver les molécules d'ADN ou les données informatiques de l'ADN ?

Dominique D. Joly, Guy Perriere

22. Crise du stockage : conserver les molécules d'ADN ou les données informatiques de l'ADN ?

Dominique Joly, Evolution, Génomes, Comportement et Ecologie (EGCE, Gif-sur-Yvette) et Guy Perrière, Biologie Biométrie Evolutives (LBBE, Lyon)

Le séquençage ADN à haut débit pose la question du volume de mémoire nécessaire au stockage des données et de leur conservation. Des besoins urgents de nouveaux développements en capacité de calcul, de stockage et d'archivage nécessitent un changement des pratiques du traitement des données.

Grâce aux méthodes omiques, les analyses à haut débit les molécules du vivant (ADN, ARN, protéines et métabolites) dans des échantillons environnementaux et humains ont engendré une forte augmentation des besoins de traitement, de stockage et d'archivage des données.

En 10 ans, la grande campagne océanographique Tara Océans a collecté 60 000 échantillons d'eau de mer, généré 1 milliard de codes-barres de taxons inconnus, identifié 100 000 nouvelles espèces de protistes et 100 millions de gènes. Le projet international *Earth Microbiome* visant à connaitre la diversité microbienne a produit, à partir des premiers 135 échantillons analysés, environ 28,8 milliards de séquences, soit environ 10 Téra-bytes (10^{13} bytes) de données et 23,3 milliards de séquences protéiques prédites. Le programme international de médecine personnalisée (*IC PerMed*), qui permettra d'améliorer le diagnostic des maladies rares, prévoit de séquencer 100 000 génomes humains, donc de produire plusieurs dizaines de péta-octets (10^{15} octets) de données par an.

Tous ces grands projets produisent des océans de séquences à gérer, à assembler et à comparer. En plus des données résultant du séquençage ADN, il est nécessaire de conserver l'ensemble des métadonnées qui caractérisent l'échantillon (données spatio-temporelle, physico-chimiques, sociologiques …). Alors que le coût de séquençage ADN a diminué d'un facteur 10^5 en 20 ans (100 millions de dollars pour un génome en 2001 contre 1000 dollars aujourd'hui), les capacités informatiques n'ont doublé que tous les 2 ans. Que faire ?

Un compromis est nécessaire entre l'intérêt de stocker les données et la récupération de l'espace mémoire pour stocker de nouvelles données. Une première approche est le tri des données pour éliminer les données inutiles et non exploitables. Une autre approche est de développer le *cloud computing* en coordination avec les systèmes de calcul haute performance bien connu du monde de la physique. Il s'agit d'améliorer la performance en termes de calcul et d'entrées/sorties des données, ainsi que d'adaptation aux variations de la demande. *GenBank*, la principale banque de stockage mondiale de données génomique, utilise déjà cette approche. Reste le problème de la bande passante pour accéder au *cloud*, qui est souvent limitante.

La conservation des molécules d'ADN au lieu des séquences informatique est également une alternative. Elle a l'avantage de s'affranchir des technologies de séquençage qui évoluent rapidement. L'ADN est communément stocké au froid entre -20°C et -80°C. Le Genopôle d'Evry a élaboré une méthode alternative dans des mini-capsules métalliques et inoxydables. Cette innovation permet de conserver indéfiniment tout type d'ADN (humain,

animal, végétal, microorganismes) à température ambiante et à l'abri des facteurs d'altération, sous une forme compatible avec un séquençage ultérieur.

L'entrée de l'ADN dans la révolution du *Big Data* induit un changement inédit de pratiques technologiques et scientifiques.

Des plaques de mini-capsules permettant de conserver de l'ADN et de l'ARN, de façon durable, à température ambiante. Photographie : www.imagene.eu.