Introduction à la modélisation en biologie

Magali Berland *

Table des matières

*magali.berland@uvsq.fr

Ce cours utilisera R et RStudio (ainsi que les packages `tidyverse`, `ade4`, `caret`, `mixOmics`.
Site web du professeur : `http://www.eldarsoft.com:10080/explore/repos`.

1 Rappels de biologie

1.1 Théorie fondamentale de la biologie moléculaire

L'ADN dirige sa propre réplication en ADN identique ainsi que sa transcription en ARN, pouvant ou non être traduit en protéines.

L'ADN est le support stable et transmissible de l'information génétique. L'ARN a une durée de vie courte et ne permet que la transmission de l'information de l'ADN vers d'autres sites (ARNm). Par la suite, le ribosome traduit l'ARN en protéines.

L'ARN et l'ADn sont très similaires, seuls les bases utilisées diffèrent (ACUT au lieu d'ATCG). Entre un humain et un chimpanzé, environ 98% des gènes sont identiques, mais leur expression est plus ou moins inhibée.

1.2 Histoire du séquençage

Le séquençage constitue en le décodage de la suite des bases de l'ADN d'individus. Il commence en 1980, où l'idée est simplement de produire des données. La base de données *GeneBank* (1982) a vu depuis sa taille doubler tous les 18 mois. En 1986, le NIH (National Insitiute of Health) est crée.

En 1989, le premier projet de séquençage du génome humain est lancé.

Un des principaux outil d'analyse de séquence est *BLAS* (Basic Local Alignment Tool). Les premières cartes du génomes humains sont publkiées par le *Généthon* (organisme français !). On pensait alors que le séquençage serait la clef pour guérir absolument toutes les maladies. Au milieu des années 1990 ont été crées les premiers séquenceurs automatisés, en parallèle de l'évolution du web qui permet des échanges plus faciles (même si il est encore plus rapide de demander un CD par la poste...). En 1997, le Centre National de Séquençage est créée. En 1998, une fondation aux Etats-Unis nommée Celera Genomics par Craig Venter (140 millions de dollars de matériel), avec pour but de finir le séquençage du génome humain, avec pour but de réqliser du profit sur d'éventuels brevets. Le coup de pression est mis et une course à la séquence se lance. Elle se termine le 26 juin 2000 par une annonce très politisée, la fin du séquençage du génome humain, déclaré patrimoine universel de l'humanité, avec impossibilité de d'en breveter une séquence. Par contre, il est possible de breveter les systèmes de détection de certaines séquences. La vraie fin du séquençage du génome est plutôt terminée en 2003. En 2007, des séquenceurs de nouvelle générations voient le jour, qui permettent de séquencer deux fois plus vite pour trois fois moins cher. Aujourd'hui, on arrive à un génome coûtant environ 1000 $ à produire et quelques heures. On espère à l'avenir réduire ce cout vers 200-300$ (environ le coût d'un examen hospitalier classique), ce qui lacerait l'ère de la médecine personnalisée (tel médicament fonctionnera ou non sur vous à cause de tel ou tel gène).

1.3 L'ère des *omics*

De nombreux domaines ont émergés de ces découvertes, dont :

— La génomique : étude du fonctionnement d'un organisme d'un organe, d'un cancer, etc. à l'échelle du génome).

Pour donner un ordre d'idée, le génome humain comporte 3 400 millions de nucléotides, pour 25 000 gènes. Les espèces ayant le plus de gènes sont les végétaux (le séquençage du blé sera publié vers cet année), ce qui est principalement dû à leur absence de mobilité.

— La transcriptonique : l'étude de l'ensemble des ARNm produit lors de la transcription d'un génome.

On y fait notamment de la quantification des ARNm (taux de transcription des gènes dans différentes conditions). Cela est très difficile car l'ARNm est très sensible, sa quantité varie énormément avec l'environnement, ce qui rend difficile l'explication des résultats. On utilisait des puces à ADN (petites

molécules d'ADN) pour réaliser des mesures, maintenant l'ARN est séquencé selon une autre méthode plus complète.

— La protéomique : l'étude des protéines d'une cellule, d'un tissu, d'un organe ou d'un organisme.
Les principaux domaines d'étude sont la quantification des protéines, l'étude de leurs interactions, leur identification et caractérisation par spectrométrie de masse. Également, la prédiction de structure s'est développée par la bioinformatique (le seul moyen exact de découvrir la structure tridimensionnelle d'un protéine passe par une analyse cristallographique, et nécessite donc la protéine pure... pas facile !).
Par exemple, la maladie de la vache folle (ou tremblante du mouton, ou Creutzfeldt-Jakob chez l'homme) provient d'une mauvaise conformation d'une protéine responsable de la création d'autres protéines dans le cerveau, générant des problèmes graves plusieurs dizaines d'années après la première mauvaise conformation.

1.4 La métagénomique

La métagénomique constitue en l'étude des micro-organismes vivant dans un certain milieu, par le séquencage de bactéries impossibles à différencier les unes des autres. On cherche alors des affiliations taxologiques (quel gène correspond à quelle bactérie), et des affiliations fonctionnelles. Un des exemple est l'étude du microbiote intestinal [1].

2 Seconde introduction à la biologie

Les biotechnologies sont en fait utilisées depuis de nombreux siècles (yahourts, fromages, saucisson, choucroute, ...). Cependant, l'ADN n'est connu que depuis 1966, et est encore un des support d'information les plus denses de nos jours. (1000 fois plus que les technologies actuelles sous forme de silicium !). Il est très robuste, contrairement à l'ARN (cf ci-dessus). L'ADN est flexible, en fonction de l'eau l'entourant, le nombre de bases par tour d'hélice peut varier de 9 à 12. Le code génétique est une transcription d'un triplet de bases de l'ADN vers un acide aminé, universelle chez tous les être vivants. On peut remarquer que ce n'est pas une bijection : il est *redondant*. On n'en est pas encore sûr, mais il pourrait que ces redondances soient en fait vecteurs de plus d'informations, encore non découvertes de nos jours.

Les phages sont des virus contenant des morceaux d'ADN qui, ayant infecté une cellule, peu se réactiver pour détruire son hote (par exemple via des UV).

Le *phénotype* est l'ensemble des caractères physiques observables d'un individu. Il correspond à l'effet ou non de l'environnement qui révèle certaines caractéristiques.

Le langage Go

Il est plus expressif que le C, ne contient pas les exceptions, intègre en natif les mêmes idées que MPI. Il est créé en 2007 par les anciens de C qui n'ont pas aimé le C++. En 2016, IBM lance Go sur leurs mainframes.

Le langage Go est postfixé, gère l'UTF-8 ; les variables et les constantes. Le mot-clef *iota* est réservé au compilateur, il augmente de 1 à chaque utilisation. Il y a un ramasse-miette et pas d'algèbre de pointeurs. toutes les variables sont initialisées par défaut. Il y a cependant des tranches de tableaux ($\rightarrow$ cool sur les GPU). Il y a deux mots-clef d'allocation mémoire : `make` et `new`, suivant que l'on veut une instance ou une référence. Le Go a une analyse d'échappement, qui permet de choisir automatiquement entre le tas et la pile.

Il existe un destructeur local : `defer`, ce qui permet de faire du code propre (une sorte de `try..catch..`). Il supporte la vectorisation dès que possible (inline). Il n'y a d'héritage, les interfaces doivent être compatible avec les objets utilisés (usine à composants ou *duck typing*).

Référence :

— The Go Programming Blueprint, Mat Ryer

1. domaine d'étude de notre professeure