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

Référence :

— The Go Programming Blueprint, Mat Ryer

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*).

1. domaine d'étude de notre professeure

2.1 Les biotechnologies

Les bases de données bioinformatiques ont évolués en suivant la loi de Moore ($\rightarrow$ des ordinateurs sont utilisées pour explorer les données. Idem pour le coût de la mégabase brut ; qui a par ailleurs dépassé la loi de Moore suite à une nouvelle méthode de séquençage.

Un gène commence par un activateur ATG et se termine par un terminateur TAA, TAG ou TGA ? Un *régulon* est composé d'un motif plus les gènes, le motif gérant la production ou non de la protéine associée au gène en fonction de l'environnement. On peut construire des expérience permettant de tester l'influence ou non de la régulation sur le gène.

Le biologiste va fabriquer des gènes modifiés et observer par électrophorèse un résultat.

Le séquençage se fait de nos jour par une amplification du signal (approche *Shotgun*) consistant à casser en petits morceaux puis utiliser des bactéries pour les multiplier ($\rightarrow$ biais de sélection). Deux mécanismes sont possible pour lire l'ADN ainsi produit : la polymérase et la ligase, modifiées pour émettre un signal à chaque base lue. Cette technique est nommée *NGS : Next Generation Sequencing*. On lit en parallèle les couleurs des séquences pour obtenir les bases des morceaux de séquences d'ADN.

Il y a quatre grandes façon de la NGS :

— 454 qui utilise de l'ATB (100 Mb *redondant* par expérience)

— SOLiD (6 000 Mb)

— Solexa GA (utiliser un ADN "hameçon" sur un substrat pour ensuite amplifier et lire les couleurs par un appareil photo.

— Helioscope

— Pacific Biosciences

Le NCBI (National center for biotechnology information) a crée GeneBank en 1992, créant alors un format de fichier .gbk (peu pratique à manipuler car non créé par des informaticiens..).

Il existe des chromosomes linéaires ou circulaires, que l'on représente souvent en GC% (proportion de GC par rapport au AC).

Definition 1 (Complexité de Kolmogorov)**.** *On définit $K(s)$ pour une suite binaire finie s, K est la taille du plus court programme qui produit s.*

On peut utiliser cette mesure pour avoir un ordre d'idée de la complexité d'un gène. Par exemple, la variabilité d'un être humain se décrit en environ 20 Mo.

3 Analyse en composantes principales

Le principes est de créer des nouvelles variables résumant les informations à partir de combinaisons linéaires des variables pré-existantes afin de minimiser la perte d'information.

4 Comparaison de séquences

L'idée intuitive est d'utiliser une vision *dot-plot* : une matrice montrant les nucléotides communes. On peut également utiliser la distance de Levenshtein, basée sur le nombre de modifications (substitution, insertion, délétion).

5 Annexe

5.1 R

La fonction `gather (key, value, ...)` permet de dupliquer les colonnes non présentes dans `...` afin d'obtenir une ligne par couple (variable, value) des colonnes sélectionné. Le nom de la colonne est mise dans

une nouvelle colonne de nom `key` et sa variable dans `value`. On peut utiliser l'opérateur `col1:col2` afin de sélectionner différentes colonnes consécutives ; et également l'opérateur `-col` pour sélectionner toutes les colonnes sauf `col`. Pour trier, on peut utiliser `arrange`.

La fonction `spread` fait l'inverse de `gather`. `separate` permet de séparer une colonne en plusieurs colonne, son contraire est `unite`. Les fonctions `filter`, `distinct` et `slice` sont également très utiles, tout comme `select` (`filter` sur les colonnes). On peut opérer sur les chaines de caractères avec `starts_with`, `ends_with`, `matches` et `num_range`.

Les données peuvent être groupée avec `group_by`, cette opération est *silencieuse*, c'est à dire qu'on ne voit pas les changements sur les tableau de données.

6 Data mining et apprentissage

Le data minng ou fouille de donnée est l recherche de *pépites d'information* pour aider à la prédiction. On utilise des techniques statistiques d'apprentissage machine pour des données de grandes dimension (big data).

En biologie, on peut utiliser des modèles de compétition proies-prédateur (l'avantage est de ne pas utiliser de variables aléatoires).

Pour rechercher des gênes, on peut utiliser des outils de machine learning pour détecter des probabilités d'occurrence de maladie ou de caractères phénotypiques.

Un modèle statistique classique est le modèle linéaire :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon$$

Avec ϵ les résidus, qui doivent être de moyenne nul si la régression est valide.

Les méthodes d'apprentissage se décomposent en supervisé (on connait à l'avance des étiquettes) et non supervisé (on recherche des classes).

On cherche un modèle *parcimonieux* : il faut éviter les modèles trop complexes (sur apprentissage) ou trop simple (sous apprentissage).

Pour mesurer la validité intrinsèque d'un modèle, on cherche a minimiser les faux positif et les faux négatifs. La sensibilité d'un modèle mesure sa capacité à donner un résultat positif lorsqu'une hypothèse est vérifiée : $\frac{VP}{VP+FN}$. L'inverse est appelée *spécificité* : $\frac{VN}{VN+FP}$ (VN = vrai négatif, FP = faux positif).

L'apprentissage est constitué des plusieurs étapes :

1. Extraction des données avec ou sans échantillonnage (sondage)
2. Exploration de données
3. Partition aléatoire (apprentissage, validation, test)
4. Pour chacune des méthodes considérées, il faut estimer le modèle pour une valeur donnée d'un paramètre puis réitérer pour tous les paramètres
5. Choisir la méthode en fonction de l'interprétabilité, de la robustesse et de ses capacités de prédiction.
6. Ré-estimer le modèle sur l'ensemble des données
7. Exploiter le modèle sur la base de donnée entière

En TD, nous avons utilisé le jeu de donnée Prostate du package lasso2.

7 Métagénomique

On a découvert la métagénomique en 2005, mais il a fallut attendre 2009 pour avoir une application : le microbiote.

L'assemblage consiste à reconstruire l'information génétique à partir de petites chaines d'ADN issues de l'amplification. Le score N50 représente est le minorant pour estimer la taille minimale des contig nécessaires pour couvrir 50% du génome. On peut vérifier l'assemblage obtenu par un dot-plot.

7.1 ggplot2

S'utilise de la manière suivante :

```
ggplot(data = DATA + GEOM_FUNCTION(mapping = aes(mapping)
```

Il existe plusieurs calques dans ggplot, qui s'ajoutent dans l'ordre suivant :

— Data (fond du calque)

— Aesthtics (légende, grille)

— Geometrics (forme de représentation)

— Facet (réaliser un tableau de plusieurs graphes)

— Statistics (modèle ou représentation statistique des données)

— Coordinates (passer en polaire, en log, etc)

— Scale (passer en linéare, log, à l'envers, etc)

— Theme (arrière-plan)

La cheat sheet dans RStudio est très utile pour retrouver rapidement ces commandes.