Le projet **ABDC**

avec la participation d'au moins (par ordre alphabétique) :

A	Matthieu	BARRET	Etienne	BELIN	**Automated**
B	Tristan	BOUREAU	Jérôme	BOURSIER	**Big**
D	Benoit	DA MOTA	Gilles	HUNAULT	**Data**
C	Emmanuel	JASPARD	David	LESAINT	**Clustering**

Au menu :

1. **A pour AlphabetS et Automated**
2. **B pour Big et Binary**
3. **D pour Data, Distances et Descripteurs**
4. **C pour Classes, Clustering, Cloud et Computers**
5. **S pour séquences de chaines de caractères (StringS)**
6. **P comme projet (enfin !)**
7. **N comme Non traité pour l'iNstaNt**

1. A pour AlphabetS et Automated

Formellement, un **A**lphabet $\mathcal{A}$ est juste un ensemble de n symboles (ou "lettres") $a_i : \mathcal{A} = \{a_1, a_2, a_3...a_n\}$. Pratiquement, la taille n de l'alphabet (son nombre de symboles en tout) a un impact non négligeable sur la performance et la finesse de la description des "mots" ou "chaines" qui utilisent l'alphabet.

Exemples :

XMP-1 pour des séquences d'ADN, $\mathcal{A} = \{ A, T, G, C \}$ et $n = 4$, soit, au mieux, deux bits de stockage.

XMP-2 pour des séquences d'acides aminés, $\mathcal{A} = \{ A, C, D...Y \}$ et $n = 20$, soit au mieux, cinq bits de stockage.

XMP-3 pour des textes "littéraires", $\mathcal{A}$ est l'ensemble des mots de la langue utilisée, soit $n > 10^4$ ou l'ensemble toutes les "formes graphiques" (comme chanter, chanté, chantée, chantait, chantaient...) soit $n > 10^5$. Astuce de stockage (sur 9 bits ?) : arbre de préfixes, de suffixes...

On peut penser à utiliser plusieurs alphabets afin de "naviguer" entre performance et précision.

Par exemple pour les séquences d'ADN, un alphabet réduit pourrait être $\mathcal{A}^* = \{Y, U\}$ et $n = 2$ avec un seul bit de stockage, si Y correspond aux pyrimidines (T,C) et si U correspond aux purines (A,G).

Pour les séquences d'acides aminés, les nombreux alphabets de propriétés comme celui de la polarité ou de la charge sont des possibilités intéressantes :

$$\mathcal{A}_P = \{"polaire","nonpolaire"\}\,;\, n = 2.$$

$$\mathcal{A}_C = \{"neutre","positive",""négative"\}\,;\, n = 3.$$

De nombreux autres alphabets (plusieurs centaines) peuvent aussi être ajoutés par discrétisation de propriétés physico-chimiques quantitatives (*FoldIndex, Hydrophobicité, Hydrophylicité...*).

Pour les textes, la lemmatisation (passage à la "racine" des mots) et la suppression des articles et autres mots de liaison mène à des alphabets simplifiés. La catégorisation (en article, nom, verbe, adverbe...) aboutit à un alphabet moins riche de possibilités mais plus "fort" en terme de sémantique et de fonctionnement.

Le terme *Automated* signifie ici qu'il faut imaginer une architecture de type "pipeline" ou "workflow" dans lequel tout est automatique : on doit juste spécifier les données, les traitements et hop (!) tout se fait par script, programme, service... sur les serveurs, dans le *cloud*, à distance ou en local...

La partie difficile est ici la vérification ou l'appréciation de la justesse des résultats puisque les volumes de données interdisent de tester "à la main" ce qu'on obtient...

La partie "intelligente" du travail est de choisir les alphabets au vu de leur intérêt réel pour la biologie, d'en imaginer d'autres "qui aient du sens".

2. B pour Big et Binary

A partir d'un alphabet $\mathcal{A}$ de n symboles a_i on construit p séquences de longueur variable : $\mathcal{S} = \{s_1, s_2, s_3 ... s_p\}$ où chaque $s_j = a_{j_1}, a_{j_2} ... a_{j_l}$. Le nombre de séquences peut être important ($p > 10^6$) et la longueur des séquences peut être très conséquent aussi ($l > 3\ 000$ Mpb pour *Homo sapiens*).

Exemples :

XMP-1 au niveau des génomes (séquences d'ADN), le site MG-RAST (initié en 2007) héberge plus de 150 000 métagénomes soit un peu plus de 59 Tbp pour en gros 792 milliards de séquences.

XMP-2 pour des protéines (séquences d'acides aminés), au 29 octobre 2014, UniProtKB/Swiss-Prot contient environ 547 000 séquences de longueur moyenne 355 aa et dont la plus longue séquence comporte 35 213 aa. GenBank, hébergé au NCBI comporte environ 178 millions de séquences. A Cambridge, l'EMBL-EBI dispose aujourd'hui de plus de 40 Po (Peta-octets) de stockage et ce ne sera peut-être pas suffisant !

XMP-3 pour des textes "littéraires", le COHA (Corpus of Historical American English) avec ses 400 millions de mots est très petit à coté de GBSAE (Google Books Standard American English) et ses 155 milliards de mots.

Pour accéder et pour traiter rapidement les données, le "standard" Fasta et .TXT est inadapté. Des éditeurs de texte **B**inaire/compacté seraient utiles, avec les outils qui vont avec... Le stockage et le codage des séquences n'est pas un détail : avec un codage inapproprié, le temps de calcul est prohibitif. Avec un stockage inadapté, les séquences ne tiennent pas en mémoire...

Exemple de réduction de données pour le fichier silva.v4.fasta : 29912 lignes, 14956 séquences, 192 Mo de texte, une fois zippé : 1,5 Mo.
...mais bien sûr la technique de codage utilisée par ZIP pour ce fichier est sans doute inutilisable pour un autre fichier.

3. D pour Data, Distances et Descripteurs

Le terme anglais **D***ata* signifie à la fois donnée (au singulier) et données (au pluriel). C'est dans le second sens qu'il faut envisager *Data*. Les individus (au sens statistique du terme) ont plein de données. Il ne s'agit pas de simples séquences mais de représentations et de projections d'informations. Un être humain ne se réduit pas à un nom et un prénom, ou à un numéro de sécurité sociale.

De même, une bactérie, surtout phytopathogène et encore plus de quarantaine n'est pas qu'une séquence d'ADN. Sa taxonomie, celle de l'hôte ou des hôtes, sa répartition géographique, son action sur les tissus etc. sont autant d'informations qu'il faut prendre en compte.

De façon similaire, une protéine n'est pas seulement une séquence d'acides aminés. C'est beaucoup plus une structure vivante, parfois localisée, qui participe à une chaine de réactions (chimiques, métaboliques...), qui s'exprime plus ou moins en fonction de conditions locales thermiques, hydriques...

Un texte "littéraire" n'est pas une simple suite de mots. C'est aussi une histoire, un récit, un ensemble de personnages, un style, une ambiance...

Une distance, au sens commun du terme, exprime la proximité ou l'éloignement entre deux objets. Mathématiquement, on peut raffiner, avec la notion d'écart, de dissimilarité, de distances particulières (métriques, ultramétriques...).

Un écueil à éviter, quand on veut comparer ou mettre en regard deux ensembles de séquences, c'est de comparer tous les x_i d'un ensemble avec tous les y_j d'un autre car cela "explose" en termes de nombres de calculs. Il faut sans doute établir des groupes-résumés de x_i et des groupes-résumés de y_j et comparer les "centres" des résumés entre eux, avant de repasser à des comparaisons plus fines à l'intérieur d'un groupe.

Le **D** de descripteur est là pour cela. Un descripteur n'est jamais qu'une fonction appliquée à une séquence, dans le but de la "décrire". On peut imaginer des descripteurs globaux, locaux, binaires, qualitatifs, quantitatifs...

Pour filtrer, trier, regrouper, les descripteurs globaux invariants par permutation comme la longueur (nombre de symboles), le XYZcontent (pourcentage de lettres X, Y ou Z...) doivent être privilégiés. On parle bien de protéines "riches" en glycine (lettre G), de séquences ADN avec un fort "GCcontent" (pourcentage de G et C).

D'autres descripteurs globaux comme les comptages de lettres (invariants par permutation), comptage de couples, de triplets... (qui eux, ne sont pas invariants par permutation) sont à envisager, au même traiter que les motifs, les k-mers, les structures répétitives, signatures et autres expressions régulières particulières.

Ensuite, il faut penser aux descripteurs "1 pour 1", qui remplacent chaque lettre par une quantité (cela rejoint la notion d'alphabets multiples) et les agréger en descripteur locaux avec des fenêtres mobiles, des courbes avec pics et intensité, des descripteurs qualificatifs avec des interprétations (comme la charge, le $FoldIndex$...), le "sens" des lettres ou des "blocs" de lettres...

Une version simplifiée peut être d'utiliser des fonctions logiques qui n'associent que vrai ou faux (présent ou absent) à chaque lettre de façon à pouvoir "caractériser" les séquences.

Définitions pour un descripteur de séquence

Un descripteur de séquence est une fonction qui s'applique à une séquence. Son résultat peut être un nombre entier ou réel, une valeur boolénne, une lettre ou une autre séquence, voire même un vecteur ou un ensemble de résultats.

Si le descripteur a besoin de toutes les lettres de la séquence, il est dit global. Par exemple la longueur d'une séquence est un descripteur global.

Si le descripteur n'utilise qu'une partie de la séquence, il est local. Par exemple le GC content est local.

Un descripteur est présentiel s'il renvoie 1 pour la présence d'un motif et 0 sinon. Par exemple la présence de "AAAAA" est un descripteur présentiel utile pour savoir si on a affaire à une bactérie.

Quelques exemples de descripteurs de séquence d'ADN (facilement généralisables)

DSC-1 A une séquence on associe son pourcentage de G.

DSC-2 A une séquence on associe son pourcentage de G ou C et sa longueur.

DSC-3 A une séquence on associe ses quatre pourcentages en A, T, G et C et sa longueur.

DSC-4 A une séquence on associe sa séquence en $\{Y, U\}$ (pyrimidines et aux purines).

DSC-5 A une séquence d'ADN on associe ses 6 séquences protéiques possibles par conversion de codon (3 nucléotides) en acide aminé.

DSC-6 A une séquence d'ADN on associe sa plus longue répétition d'un même nucléotide sous la forme (lettre,longueur,position).

DSC-7 A une séquence d'ADN on associe le nombre d'occurences et la position d'une suite de motifs prédéfinis, par exemple (et dans cet ordre) les motifs TATA, ATG, TAA, AAAA.

4. C pour Classes, Clustering, Cloud et Computers

Là, les **C**hoses se **C**ompliquent ! Les classes, ce sont des groupes ou des sous-ensembles particuliers, qui rassemblent les éléments d'une même classe et les opposent aux autres classes. La littérature informatique et statistique déborde de méthodes, de concepts pour "faire" ces classes, les reconnaitre, les remplir.

Oui, mais, *quid* de la vitesse d'exécution ? Le coût d'un modèle linéaire, d'une régression logistique prohibent d'utiliser ces outils sur de très grands ensemble de données, sans parler de la précision (illusoire) des résultats. Vouloir déterminer des intervalles de confiance par rééchantillonnage est encore plus interdit. Que reste-t-il alors comme méthodes rapides et fiables ?

Faut-il stocker de gigantesques matrices de distances afin de ne pas tout recalculer en fonction de nouvelles données ?

Et surtout, quel est le lien avec les séquences ? Quel rapport entre, disons, la distance de Jaccard et les séquences d'ADN ? Ou entre une distance ultramétrique et des courtes séquences d'acides aminés ?

C'est pourquoi il faut esayer de définir des descripteurs de classe ayant "un sens".

Un **descripteur de classe** est un ensemble de descripteurs de séquences dont on fixe les valeurs. Un "bon" descripteur de classe est un descripteur qui permet de dire si une séquence peut appartenir à la classe. Un descripteur caractéristique de classe est un descripteur qui garantit que la séquence appartient à la classe. En d'autres termes, il s'agit *mathématiquement*, d'une fonction **injective** sur les classes, comme un mot de passe l'est pour garantir un utilisateur.

Exemple naif :

On construit un descripteur d'architecture de classe qui regroupe

- un descripteur de séquences lié à un coefficient numérique exprimant "l'ordre" ou "la régularité" dans les séquences,
- un descripteur de séquences basé sur la longueur de la séquence,
- un descripteur de séquences basé sur la plus grande répétition dans la séquence,
- un descripteur de séquences basé sur des motifs spécifiques,
- un descripteur de séquences qui utilise la taxonomie liée à la séquence.

5. S pour séquences de chaines de caractères (StringS)

Les objets de base manipulés par les statistiques sont les vecteurs et les "vrais" tableaux rectangulaires de données numériques quantitatives. Les données sous formes de **S**équences présentent de vrais enjeux conceptuels, de nouvelles méthodes à inventer et à tester, des algorithmes à imaginer.

En particulier les notions d'homologie, d'alignement, de distance mettent en jeu des notions différentes des calculs "classiques" sur vecteurs et matrices. La notion de descripteur local, de fenêtre mobile, de signature contextuelle sont typiques de telles différences.

De plus, le contenu des vecteurs qui ne sont pas directement numériques mais des chaines de caractères avec des interprétations possibles induisent des calculs différents. Pour simplifier, au lieu de $V \in \mathbb{R}^n$, on travaille plutôt avec $V \in \{A, T, G, C\}^n$ ou $V \in [\![1, 20]\!]^n$ si $[\![a, b]\!] = [a, b] \cap \mathbb{N}$.

C'est pourquoi il faut penser à de nouvelles méthodes, pas juste des calculs mais aussi des interprétations, des caractérisations... car on dispose de données **riches** en termes d'explications biologiques, contrairement aux simples et "pauvres" valeurs de nombres réels sans dimension et sans unité.

6. P comme projet (enfin !)

Notre projet, modeste mais réaliste, est de constituer un groupe de travail en bioinformatique spécialisé en "classification caractérisante" (supervisée ou non) de gros volumes de données. Dans un premier temps, il s'agit de se faire "parrainer" et reconnaitre par les "grands partenaires" sur un thème proche du *binning* en métagénomique (données ADN) mais les concepts et les outils doivent à moyen terme pouvoir s'appliquer à des séquences protéiques.

Il parait raisonnable de se limiter pour l'instant au microbiome humain des intestins et au microbiome des graines et semences pour rester dans les "lignes de recherche" de nos laboratoires mais le formalisme et les concepts pourraient permettre de classer et caractériser d'autres séquences.

En deux mots : on veut retrouver (et quantifier) dans un ensemble U de séquences d'étude y_i des caractéristiques issues d'un ensemble structuré K de séquences x_i de référence.

Au lieu de comparer "bêtement" tous les y_i avec tous les x_i, on commence par regrouper les y_i par des techniques similaires à celles qui ont permis de classer les x_i en classes de référence C_k.

Dans le projet, on construit des bases de données de descripteurs de classes et on associe aux séquences les valeurs des descripteurs (on ne recopie donc que l'identifiant de la séquence des grandes bases de données de séquences).

On compare ensuite les valeurs des descripteurs de classes connues avec celles obtenues pour les classes inconnues afin de détecter les classes probables. Ensuite, on revient à des comparaisons d'une séquence inconnue $versus$ une classe connue pour

- soit considérer qu'on a trouvé la classe de la séquence
- soit reconnaitre qu'on a une nouvelle classe
  (et donc des nouveaux paramètres de descripteurs de classe)

Ce qu'on veut faire, ce n'est donc pas comparer, assembler, classifier, séquencer car d'autres grands laboratoires et d'autres équipes le font (et certainement mieux que nous ne le ferions), mais construire une base de données de descripteurs de référence pour détecter, compter, reconnaitre des gènes, des bactéries, ou, de façon générale, des classes de séquences.

On mettrait ainsi à disposition des descripteurs et des outils pour construire et tester de nouveaux descripteurs. De plus ces decripteurs devraient pouvoir être "parlants biologiqumeent" au lieu de se définir comme un seuil de e-value pour blast, par exemple.

Ce qu'il nous faut comme moyens humains, matériels et financiers :

1. des post-docs pour nous aider à concrétiser nos idées,
2. des Tera-octets et des heures de calculs pour les tester et les réaliser,
3. des partenaires industriels européens,
4. des financements importants pour les déplacements et de la formation en doubles voire en triples compétences, pour du matériel performant et pour monter des prototypes,
5. des collaborateurs "académiques" pour nous aider à démarrer.

Quelques pistes :

1. IRHS, LERIA et HIFI
2. la plate-forme GenOuest à Rennes
3. via C. Manceau, ANSES (Agence nationale de sécurité sanitaire...)
4. projet national, ANR...
5. MIG (Mathématique, Informatique et Génome) INRA Jouy en Josas

7. N comme Non traité pour l'iNstaNt

Ce à quoi vous avez échappé (pour cette fois)

XMP-4 des séquences musicales comme des partitions transcrites via l'alphabet musical { A (la), B (si), C (do), D (ré), E (mi), F (fa) et G (sol)... } hérité de la Grèce antique.

XMP-5 des séquences de "textes oraux" transcrits selon l'API (alphabet phonétique international).

XMP-6 des séquences de textes pour détecter du plagiat, de la fraude.

XMP-7 des séquences (?) correspondant à des images sans bruit de fonds.