Le projet **ABDC** : les descripteurs

Au menu :

1. **Philosophie du projet**
2. **Vision globale des descripteurs de séquences**
3. **Exemples de descripteurs de séquences**
4. **Les descripteurs en revue**
5. **Appel à contributions**

1. Philosophie du projet

ABDC doit permettre de classifier (regrouper) et de classer (associer à une classe connue) des séquences nucléotidiques sans passer forcément pas des alignements comme point d'entrée.

Une première partie consiste à recenser et mettre en ligne les caractéristiques des génomes, ce que nous nommerons *descripteurs*. Une base de données de descripteurs avec des résultats pour les génomes bactériens déjà complètement assemblés en est une partie importante, accessible – comme preuve de concept – via la page Web EODD (*Encyclopedia Of DNA Descriptors*) :

`http://forge.info.univ-angers.fr/~gh/Abdc/eodd.php`

La seconde partie consiste à proposer des méthodes de classification rapides (pour ces "Big Data") à l'aide de comparaisons entre descripteurs, c'est-à-dire utiliser des calculs de distances non pas sur les séquences mais sur des objets issus de génomes comme par exemple les images CGR (*Chaos Game Representation*) associées.

Il s'agit donc de distances *indirectes* et non pas de distances *directes* sur les séquences.

2. Vision globale des descripteurs généraux de séquences

Un descripteur est **un peu** comme une fonction au sens mathématique du terme, avec un élément et son image, ou au sens informatique du terme une fonction qui passe d'une entrée à une sortie. Exemples :

f : $x \rightarrow pI(x)$ # point isoélectrique

`nombreDeA` = fonction(chaine) : renvoie(entier) # nombre de fois où on voit A

La différence ici est au niveau des entrées, des sorties et des temps de calcul.

Pour les entrées :

- x est une séquence ou un ensemble de séquences, pas un vecteur numérique.

- un génome n'est pas qu'une chaine de caractères (objet, attributs multiples).

- selon le contexte, `A` représente le nucléotide Adénine, l'acide aminé Alanine ou la lettre A dans une langue donnée.

Pour les sorties :

- le nombre de A n'est peut-être pas pertinent (son pourcentage est peut-être mieux).

- le nombre de A n'est peut-être pas caractéristique suivant le type d'entrée (un *read*, un génome, des génomes, une séquence protéique, des séquence protéiques, un texte de M. Proust...).

- ce n'est peut-être pas A seul qu'il faut envisager (TATA plutôt, par exemple).

Pour les temps de calculs :

- une séquence protéique fait quelques centaines d'acides aminés en moyenne comme en médiane.

- le génome de Pseudomonas putida GB-1 comporte sans doute 6 078 430 nucléotides (6 millions !) dans une archive zip de 1,8 Mb pour un volume de 5,9 Mb une fois décompressé.

- une comparaison lettre à lettre de deux génomes de 10^6 nucléotides en programmation dynamique passe par l'utilisation d'une matrice $10^6 \times 10^6$.

3. Exemples de descripteurs de séquences

Voici un descripteur qui prend en entrée une séquence qui est ici une chaine de caractères et dont la sortie est multiple et hétérogène en type.

Exemple 1 : un descripteur générique (*càd* qui s'applique à tous types de chaines)

lngstCstSubstr(`"TATACCCG"`) $\longrightarrow$ (3,`"C"`,5)

Ici, lngstCstSubst vient chercher dans une chaine de caractères la plus grande sous-chaine de caractères constante. Par exemple pour `TATACCCG` c'est `CCC` de longueur 3 vue à partir de la position 5, soit **(3,"C",5)** qui se lit 3 fois C à partir de la position 5.

Intérêt pour la linguistique humaine en littérature : sans doute aucun.

Intérêt biologique : il y a de fortes chances pour qu'une très longue chaine constante fournisse une indication de mauvaise qualité de séquençage pour une bactérie.

Exemple 2 : un descripteur nucléotidique

Le descripteur lngstRepSubstrGt3 vient détecter dans une chaine de caractères la plus grande sous-chaine répétée de longueur > 3.

Intérêt biologique : cela correspond peut-être aux "`TATA` boxes", ou aux promoteurs ou aux "start transcription factors" ou aux gènes.

$\Longrightarrow$ Ce descripteur peut s'appliquer aux *reads* comme aux génomes.

Exemple 3 : un descripteur génomique défini *a priori* pour le genre *Pseudomonas*

lngstComSubstrPs({ génomes séquences du genre *Pseudomonas* })
= la plus grande sous-chaine commune à tous les génomes connus
des bactéries du genre *Pseudomonas*

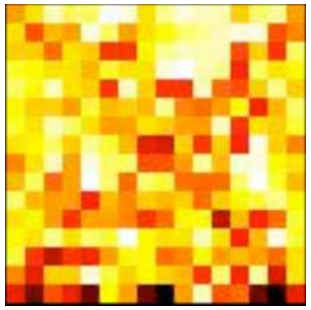
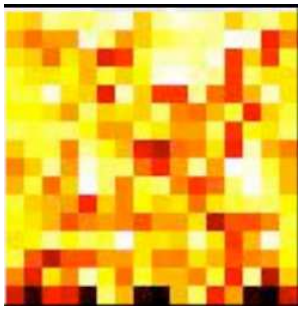
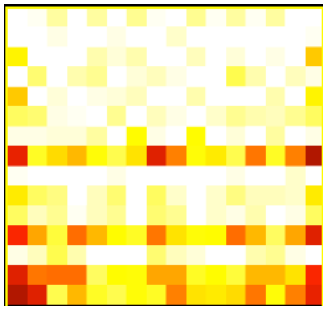
Intérêt biologique : est-ce un gène, un opéron ? Ou est-ce plus petit, plus grand ? Est-il réutilisable pour les rangs taxonomiques au-dessus (famille...phylum) ? au-dessous (espèce, pathovar...) ?

Exemple 4 : un descripteur qui renvoie une ou des images

imgGenomesCGR({ génomes séquences d'un genre donné })
= { ensemble d'images CGR }

Variante : on renvoie une "image consensus" de l'ensemble des images.

CGR signifie *Chaos Game Representation* comme ci-dessous :

		
Pseudomonas	*Pseudomonas*	*Bombyx* (pour mémoire)

4. Les descripteurs en revue

Il faut dès à présent envisager des descripteurs

- **généraux** qui s'appliquent à tout type de chaine de caractères,

- **bactériens** [...],

- **locaux** relatifs à un genre, une espèce...,

- **qui s'appliquent des descripteurs** pour évaleur leur qualité, leur pertinence...

- **...**

qu'il faudra coupler à des descriptions de ces descripteurs pour fournir du recul et de l'expertise quant à leur utilisation.

5. Appel à contributions

5.1 Choix de l'ADN bactérien

Il semble raisonnable, dans un premier temps, de se limiter à l'ADN bactérien pour les raisons suivantes :

- les génomes bactériens sont plus petits que les génomes fongiques et humains ;
- les gènes et les *ssbi* (sous-séquences biologiques d'intérêt) sont en général peu répétés ;
- presque tout le génome est "utile" ;
- la taille des *reads* actuels (quelques centaines de bases) n'est pas "trop petite" par rapport à la taille d'un gène bactérien (environ 1 kb) ;
- les collègues de l'I.N.R.A. travaillent actuellement plus sur les génomes bactériens que sur les autres génomes...

5.2 Descripteurs informatiques systématiques

Dans la mesure où les entrées (*reads*, gènes, génomes, protéines...) sont des chaines de caractères définies sur un alphabet, de nombreux descripteurs informatiques sont faciles à imaginer

- pour une chaine :
  - fréquence de chaque lettre dans la chaine, pourcentage associé ;
  - fréquence de chaque sous-séquence de n lettres dans la chaine, pourcentage associé (n=2, 3, 4... mais jusqu'où aller ?) ;
  - plus grande sous-chaine constante (avec position, longueur, nombre de répétitions) ;
  - plus grande sous-chaine répétée (avec position, longueur, nombre de répétitions)...

- pour un ensemble de chaines :
  - plus grande sous-chaine commune ;
  - plus grande sous-chaine commune constante ;
  - plus grande sous-chaine commune répétée au moins n fois (choix de n ?) de longueur entre $lngMin$ et $lngMax$;
  - plus grande sous-chaine commune caractéristique discriminante (du genre, de l'espèce etc.)...

5.3 Descripteurs mathématiques et physiques systématiques

Chaque comptage précédent peut s'intégrer dans un test statistique sous hypothèse d'équirépartition et il faut donc ajouter à chaque comptage le petit p correspondant.

On peut aussi imaginer des compléments statistiques plus sophistiqués comme des modèles de Markov (cachés ou non) des fréquences d'apparition des lettres, des sous-séquences répétées, des sous-séquences caractéristiques...

Chaque chaine peut s'analyser aussi comme un signal en théorie du signal et donner lieu au calcul de représentations comme les images CGR, être codée par ondelettes...

5.4 Descripteurs biologiques

A coté de ces descripteurs systématiques, il faut **absolument** ajouter des descripteurs connus dans la littérature et il y en a beaucoup, dont certains ne sont pas décrits de façon informatisable (et on ne les connait alors que par annotation suite à des expérimentations) :

- gènes, promoteurs, débuts de facteur de transcription dont gènes de virulence, gènes de ménage (heureusement, il n'y a pas d'intron ni d'exon dans les génomes bactériens) ;
- autres *ssbi* (sous-séquences biologiques d'intérêt) : intégrons, opérons, uber-opérons, éléments transposables...
- régions et domaines spécifiques liés au genre, à l'espèce... (ilots CPG, iles-O, iles-K, marqueurs moléculaires procaryotes bien conservés,
- ...

5.5 Contributions

Il est clair que recenser et déposer les descripteurs cités ci-dessus dans une base de données associée à un site Web demande une forte collaboration entre les participants et une grande expertise pour que la base de données des descripteurs puisse être efficace.

Il faudra non seulement mettre à disposition des moyens de calculer (sur le site ou en local) les descripteurs pour des séquences utilisateur mais aussi fournir de l'aide et du conseil sur le choix des descripteurs et la portée de leur utilisation.

Par exemple : peut-on se servir du contenu GC pour des *reads* ? ou encore : comment utiliser une image CGR pour attribuer un genre à une séquence ?

Une grande difficulté à ne pas sous-estimer est l'ampleur de la tâche : même avec seulement 1000 (mille génomes) bien assemblés en une seule séquence, il y a beaucoup de calculs (parfois longs à effectuer) et à stocker dans la base de données. Chaque nouveau génome complétement assemblé (obtenu par veille technologique) devra être sytématiquement ajouté, comparé sans avoir à "tout recalculer dans tous les sens."

5.6 *Pseudomonas et al.* pour commencer

Il faut bien commencer à tester toutes ces idées et les descripteurs associés.

Nous avons choisi avec M. Barret de nous focaliser pour cette année (dans cet ordre) sur :

1. *Pseudomonas putida* séquencé en 5 exemplaires,
2. tous les autres génomes séquencés du genre *Pseudomonas*,
3. le seul génome séquencé de *Helicobacter hepaticus* (comme "outgroup"),
4. les 10 génomes séquencés de *Helicobacter pylori*,
5. le génome de *Fusobacterium nucleatum*.

Nous sommes ouverts à tout autre choix et à toute discussion sur *Protéobactéria* (epsilon-protéobactéries et gamma-protéobactéries) versus *Firmicutes* (dont Bacillus...).

Votre aide est la bienvenue pour indiquer des descripteurs, les programmer ou tester les implémentations, rédiger les "conseils d'expert" sur les descripteurs etc.

5.7 Et donc...

MERCI

et

AU TRAVAIL tous ensemble !