Étude des séquences d'ADN

MARIA BARBI (10 JANVIER 2005)

1 Introduction

1.1 Le code génétique

La fonction de l'ADN est de fournir les indications nécessaires à la fabrication des protéines dont l'organisme a besoin. L'instruction pour cette opération est inscrite dans la molécule d'ADN sous la forme d'une séquence de quatre bases azotées, Adénine (A), Cytosine (C), Guanine (G), Tymine (T). La synthèse protéique commence par une phase de *transcription*, où une partie d'ADN est "recopié" dans une séquence d'ARN messager, puis cet ARN est utilisé pour assembler dans le bon ordre la séquence d'acides aminées qui forme la protéine (*traduction*).

Le *code génétique* est le système de correspondance qui permet à la cellule de fabriquer des protéines à partir de son ADN. Il "traduit" des "mots" de l'ADN en chaînes d'acides aminés. Ces "mots" sont formés par groupes de trois bases, appelées *codons*. Chaque codon correspond à un acide aminé (par exemple, le codon AAT de l'ADN correspond à l'acide aminé "leucine") ; l'ordre des codons correspond à l'ordre des acides aminés dans la protéine. Le code génétique est universel : chez tous les êtres vivants, un même triplet de nucléotides correspond à un même acide aminé. Le code génétique est dégénéré : les codons différents ne codent pas forcement pour des acides aminés différents. Les différents codons qui codent pour le même acide aminé sont dit synonymes. En effet, quatre nucléotides (A, C, T, G) sont possibles pour chacune des trois positions du codons : on peut ainsi former $4 \times 4 \times 4 = 64$ codons différents, alors que les acides aminés sont en nombre de 20 seulement. Il y a donc une certaine redondance dans le code génétique.

On a remarqué que tous les codons associés à un même acide aminé ne sont pas présents avec la même fréquence dans les séquences codantes. Ce biais dans l'emploi des codons est propre au génome : quand il existe, il est identique pour la majorité des protéines d'un organisme. On a aussi montré qu'il y avait une forte corrélation entre la fréquence d'utilisation des codons synonymes et la fréquence de leur ARN de transfert dans le cytoplasme (*ARNt*, les petites unités d'ARN chargées de "porter" chaque acide aminé en solution et de lire l'ARN messager pour aligner ces acides aminés avec sa séquence). Cette corrélation est importante pour les gènes hautement exprimés alors que les gènes peu exprimés montrent des fréquences des codons synonymes plus ou moins similaires. On en a déduit que les codons synonymes les plus utilises sont ceux qui sont reconnus par la classe d'ARNt la plus abondante. Dans ces PROJETS, on pourra mettre en évidence ce caractère non aléatoire de l'utilisation des codons dans certaines séquences d'ADN grâce à une étude de leur spectre en fréquence.

1.2 ADN codant et non codant

Le génome humain contient autour de 3 milliards de bases azotées. Cette information pourrait suffire à coder plus de trois millions de types de protéines en sachant qu'une protéine est formée en moyenne par 300 acides aminés (donc 900 bases ou 300 triplets). Or, cela ne correspond pas au nombre de protéines différentes que l'on

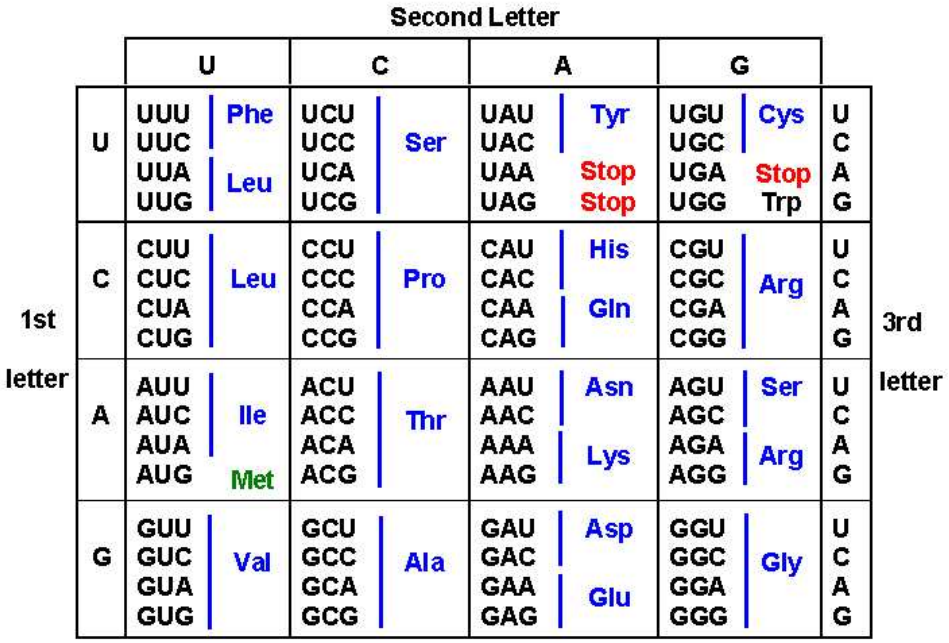

FIG. 1 – Le code génétique

peut retrouver chez les organismes supérieures : on estime qu'il y a entre 10000 et 30000 protéines différentes dans un organisme. Il y a donc un excès important de bases sur l'ADN. On parle alors d'ADN dit *codant*, et d'ADN *non codant* qui ne sert pas (pas directement en tous cas) pour la synthèse des protéines. La fonction des séquences non codantes n'est pas bien connue, mais on commence aujourd'hui à envisager soit une fonction régulatrice des phénomènes plus spécifiquement associés à la transcription et à la réplication de l'ADN, soit un rôle dans l'organisation de la molécule dans le noyau cellulaire, etc. Cet ADN non codant n'est généralement pas ou peu présent dans les procaryotes (cellules privées de noyau cellulaire) et dans les virus, mais peut être une portion très importante du génome des eucaryotes (cellules avec noyau cellulaire, organismes supérieurs). On estime que seulement 10 % des bases de l'ADN humain sont codantes.

Des régions non codantes peuvent être présentes aussi dans les gènes (unités de transcription). La région codante d'un gène est appelée *exon*. Un gène renferme plusieurs exons et ceux-ci ne sont pas disposés les uns à côté des autres. Ils sont séparés par des régions appelées *introns* dont la séquence n'est pas utilisé lors de la synthèse d'une protéine. Il peut exister plusieurs introns dans un même gène, avec un nombre total de bases plus grand que pour les exons. Les PROJETS proposent aussi de comparer le spectre d'un ADN procaryote à celui d'un ADN eucaryote, et d'en analyser et interpréter les différences sur la base des considérations précédentes.

1.3 Corrélation longue portée dans l'ADN

Si le rôle des nombreuses séquences non codantes est lié à l'organisation spatiale de l'ADN, l'ordre temporel des phénomènes qui l'impliquent, la régulation de l'intensité de ces phénomènes, leur coordination etc., on est tenté d'imaginer que ces séquences ne soient pas aléatoires, mais possèdent une organisation interne complexe. De plus, leur rôle dans l'organisation du fonctionnement de l'ADN laisse soupçonner que cette organisation soit plutôt répandue sur des longues distances.

En mots plus concrets, une corrélation doit exister entre portions de séquences bien distantes le long de la molécule.

En effet, des corrélations longue portée on été mis en évidence pour les ADN riches en séquences non codantes. Il existe différentes hypothèses sur l'origine exacte de ces corrélations : une réponse définitive n'existe pas encore, et ce problème fait l'objet d'études actuellement en développement, qui impliquent, d'ailleurs, des techniques d'analyse des signaux de plus en plus évoluées.

PROJETS 1 ET 2

Récupérer dans la banque de données *NCBI* (*National Center for Biotechnology Information*, site internet `http://www.ncbi.nlm.nih.gov/`) la séquence de bases du **génome** complet du virus bactériophage Lambda (**lambda phage**, séquence de 48502 bases) et la séquence de **nucléotides** correspondante au récepteur des cellules T humain nommée **HUMTCRADCV** (97630 bases).

Utiliser le mode de *display* nommé *FASTA* (ou *TinySeq XLM*) pour obtenir les séquences sous un format plus compact et les sauver dans deux fichiers. Éliminer tous les commentaires et modifier le format des fichiers de telle sorte que la séquence de bases soit écrite sur une colonne continue.

Écrire un programme SCILAB pour lire la séquence de lettres, puis la traduire en séquence numérique en utilisant un code de son choix (on peut par exemple associer les bases puriniques A et G à la valeur 1 et les pyrimidiques, C et T à la valeur -1, ou à 0. Alternativement, on peut associer différemment les bases en couples, ou bien en choisir une à opposer à toutes les autres, etc.).

PROJET 1

Étudier le *spectre en fréquence* des différentes séquences considérées comme des signaux temporels échantillonnés, les comparer à celui d'un signal du même type complètement aléatoire. Décrire les résultats obtenus et essayer de les expliquer. Que peut-on conclure sur la nature des deux ADN considérés ?

Étudier l'effet de l'usage d'un code différent pour la numérisation du signal.

Étudier *la fonction corrélation* des différentes séquences considérées comme des signaux temporels échantillonnés, les comparer à ce d'un signal du même type complètement aléatoire. Décrire les résultats obtenus et essayer de les expliquer. Que peut-on conclure sur la nature des deux ADN considérés ?

Étudier l'effet de l'usage d'un code différent pour la numérisation du signal.

PROJET 2

Une petite bibliographie de référence sur ces thèmes est donnée. Faire une synthèse de ces articles en expliquant quelles sont les méthodes utilisées pour l'étude des séquences génomiques, les informations que ces méthodes permettent d'obtenir et l'interprétation des résultats obtenus par les différents auteurs. En collaboration avec l'étudiant qui s'occupe du premier projet, expliquer les résultats obtenus ainsi que les difficultés et les avantages des différentes méthodes d'analyse des séquences.

Bibliographie :

Voss, "Evolution of long-range Fractal correlations and 1/f noisein DNA base sequences", Prhysical Review E **68**, 3805 (1992).

Buldyrev *et al*, "Long-range correlation properties of coding and noncoding DNA sequences : GenBank analysis", Phys. Rev. E **51**, 5084 (1995)

Allegrini *et al*, "Dynamical models for DNA sequences", Physical Review E **52**, 5281 (1995) ("Introduction").

Li, "The complexity of DNA", Complexity **3**, 33 (1997).

Note technique

Pour traduire la séquence de lettres en séquence numérique, on peut envisager de passer par la séquence numérique des codes `ascii` associés aux quatre lettres par SCILAB, puis substituer à ces valeurs les valeurs souhaitées.

Alternativement, le passage à la séquence numérique peut être fait directement avec l'éditeur de texte (dans une copie du fichier texte de départ).

Fonctions SCILAB qui peuvent résulter utiles pour le projet :

```
ascii
find
fft
strsubst
read
str2code
rand
round
corr
```

Voici enfin des instructions pour les traduire (sous scilab) en séquences numeriques.

1) scilab contient un "codage" propre pour tous les caractères :

```
-->ascii("A")
ans =
65.
```

on peux donc l'utiliser, et il est bien de le definir dans son propre fichier avec des noms :

```
// ascii code of A
ca=65 ; // code of 'A'
```

(de meme pour C G et T)

2) nous voulons par contre un code à nous, à definir aussi, par exemple :

```
my_ca=1 ; // my code of 'A'
```

(et aussi pour T G C)

3) la séquence peut se lire avec `read` : il faut specifier le nom du fichier, le nombre de lignes, le nombre de colonnes, et le format : pour ce dernier utiliser `'(a)'` .

4) Une fois obtenue la sequence (dans une variable `seq` par exemple) on peux la convertir dans le code ascii avec la commande `ascs = ascii(seq)`

(faire `clear seq` en suite pour nettoyer la memoire).

5) Apres on definira un vecteur nul (qu'on appelle ici `s`) de longuer appropriée,

6) et pour changer le code, on peut utiliser la methode suivante :

```
idx=find(ascs==ca) ;
s(idx)=my_ca*ones(length(idx),1) ;
```

(de meme pour C T G : à chaque passage verifier ce que `s` devient..)

(faire `clear ascs` après).