C’est la séance 5.2 de la formation 2015-2016.

Ce TP met en œuvre la recherche de sous-chaînes et l’utilisation d’heuristiques dans un contexte d’application à la génomique.

Il est inspiré par le document élaboré par Pierrick Bouttier au sein du groupe Algorithmique et Logique de l’IREM, lui même basé sur l’article *À la recherche de régions codantes* de François Rechenmann sur http://interstices.info, dont la lecture est par ailleurs chaudement recommandée.

Sommaire

Introduction

On reprend ici les deux premiers paragraphes de l’article de François Rechenmann, ce qui nous évitera de mal les reformuler (on le fait juste après en fait).

> *Stricto sensu, le génome d'un organisme est l'ensemble de ses gènes ; autrement dit, l'information nécessaire à ses cellules pour synthétiser les protéines qui assurent des fonctions diverses : structure, transport, catalyse, etc. Par extension, le terme génome désigne également le support physique de cette information, la molécule d'ADN (Acide DésoxyriboNucléique), composant des chromosomes présents au sein de chacune des cellules de l'organisme. L'ADN est un enchaînement de nucléotides de quatre types différents distingués par leur base azotée : adénine, thymine, cytosine et guanine, et notés par les initiales A, T, C et G. Et c'est cet enchaînement qui code l'information génétique, au même titre qu'une suite de 0 et de 1 peut coder un son, une image ou une suite d'instructions.*
>
> *Au sein d'un gène, et plus précisément au sein de sa région codante (ou CDS pour CoDing Sequence), la suite des triplets de nucléotides, appelés codons, dicte la séquence en acides aminés de la protéine. La correspondance entre les 64 (4^3) codons possibles et les 20 acides aminés constitue le code génétique, identique à peu de variantes près chez tous les organismes vivants.*

En bref : le génome d’un organisme est une suite de nucléotides notés par les lettres `'A'`, `'T'`, `'C'`, `'G'`, tandis que les gènes sont des suites de codons (triplets de nucléotides) consécutifs dans le génome.

Mais toutes les suites de triplets ne codent pas nécessairement des protéines : comment les identifier au sein du génome ?

Début et fin des séquences codantes

Citons encore :

> *Les biologistes savent que le début d'une région codante dans un génome bactérien est marqué par un triplet ATG, appelé* start*, et sa fin par l'un des triplets TAA, TAG ou TGA, appelés* stop*.*

Deux premiers problèmes se posent :

- on ne sait pas par avance dans quel *sens* il faut lire le génome pour obtenir une séquence codante ;
- les gènes ne sont pas nécessairement alignés sur des positions multiples de trois nucléotides.

Il y a donc trois *phases* possibles dans chaque direction, et six manières en tout de lire dans le génome.

De plus, rien n'empêche des codons *start* de se ballader au milieu d'une séquence codante : ils codent alors l'acide aminé correspondant. Donc si on peut deviner où s'arrêter (les codons stop ne sont jamais traduits en acides aminés), il n'est pas évident de savoir où on commence.

Pire : rien n'empêche des codons *start* ou *stop* de se ballader à l'extérieur des séquences codantes ! À partir de ces informations, on ne peut donc qu'essayer de deviner où se trouvent les séquences potentiellement codantes, au risque d'obtenir quantité de faux positifs.

C'est la première étape de ce TP.

Étant donnée une chaîne de caractères (nucléotides) `'A'`, `'C'`, `'G'` et `'T'`, on cherche à obtenir la liste des séquences potentiellement codantes, données par le sens de lecture et les indices des codons *start* et *stop*. On vous demande de réaliser l'heuristique suivante :

- inspecter toutes les sous-chaînes délimitées par deux codons *stop* en phase,
- au sein de ces sous-chaînes chercher le premier codon *start* en phase,
- ce qui se trouve entre ce codon *start* (inclus) et le codon *stop* qui le suit (non inclus) est une séquence potentiellement codante,
- ne retenir que les séquences potentiellement codantes de plus de 300 nucléotides.

Note : On ne se préoccupera pas de la quantité de mémoire gaspillée en codant chaque nucléotide comme un caractère (c'est seulement quatre fois trop).

Afin de ne pas travailler dans le vide, on utilisera les résultat du séquençage de l'ADN d'un organisme particulièrement simple : *Bacillus subtilis subsp. subtilis str. 168*. Les données sont issues de la base European Nucleotide Archive (ENA) et se présente sous forme textuelle par lignes de 60 nucléotides.

Pour référence, voici la liste des nucléotides 285000 à 290999.

```
TTTAACGACTATCGCGGATATGTCCGCTGTACAGTGACGCCTCACCAAGTGGAAAGCCGA
TTATCGGGTGATGCCATTTGTGACCGAGCCGGGCGCAGCCATTTCCACGCGGGCTTCATT
CGTTTACCAGAAAGACCAAACCGGGTTGAGAAAGGTATCATCCACAACAATCCAAGGCGG
GGTGAAGCAATCCGATGAGGTCGAAGAGGATCGTTTCTTTTCGCACAACAAAGCCCACGA
AAAACAAATGATTAAGAAGCGTGCAAAAATCACGAATTAAGGAGTGGAAATTATGTTTTC
AAACATTGGAATACCGGGCTTGATTCTCATCTTCGTCATCGCCCTCATTATTTTTGGCCC
TTCCAAGCTGCCGGAAATCGGGCGTGCCGCCGGACGGACACTGCTGGAATTTAAAAGCGC
```

```
CACAAAATCACTTGTGTCTGGTGATGAAAAAGAAGAGAAATCAGCTGAGCTGACAGCGGT
AAAGCAGGACAAAAACGCGGGCTGAATGCTGATGAGGCAGACACCGGGTCTGCCTCTTTT
TTTATGAAAGGGAGGGCTTTTTTGAATGGATAAAAAAGAAACCCATCTGATCGGGCATTT
AGAAGAGCTTCGCCGCCGGATTATCGTCACCCTTGCGGCATTTTTTCTATTTCTCATCAC
GGCTTTTTTGTTCGTACAGGACATTTATGACTGGCTGATCAGGGATTTGGATGGAAAGCT
GGCTGTGCTAGGACCGAGTGAAATCCTCTGGGTGTATATGATGCTTTCCGGCATTTGTGC
CATTGCGGCTTCTATCCCTGTTGCCGCGTACCAGCTGTGGCGTTTCGTTGCACCGGCGCT
GACTAAAACGGAGCGCAAGGTGACGCTCATGTACATCATGTACATACCAGGTTTATTTGC
GTTGTTTTTGGCGGGCATCTCCTTCGGATACTTTGTCTTGTTTCCGATCGTGCTCAGCTT
TTTGACTCATTTATCCTCCGGCCACTTTGAAACGATGTTTACGGCTGACCGCTACTTTAG
GTTTATGGTGAATTTGAGCCTGCCGTTCGGCTTCTTGTTTGAGATGCCCTTGGTGGTGAT
GTTTTTAACAAGGCTGGGCATCTTAAATCCTTACAGACTGGCCAAAGCGAGAAAGCTTTC
CTATTTTCTGCTGATTGTCGTGTCCATATTGATTACACCGCCTGATTTTATTTCTGATTT
TCTCGTGATGATCCCGCTTCTTGTCCTGTTTGAAGTGAGTGTCACCCTATCGGCGTTTGT
CTACAAAAAGAGGATGAGGGAAGAAACAGCGGCGGCCGCTTAGTGCAGCGTACCACCCGG
TGACTTCACATCCTCATCATATTGTGCGGCCGTAACAGCGGCGATTCTCAATGCCCGGAC
AATCGTGTCCAGGCTGAGGCTCGGCGCTGTTTTGTCGATTGTTTGCTGCGGAATGTAAGG
AATATGAATAAAACCGCCGCGAATGTGTGGGGATGTCCGGCTAATGTGATCCATTAACCC
GTAGAACAAATAGTTGCATACAAAGGTCCCCGCTGTGTAGGAAACCGCAGCTGGAATGCC
GTGTTCCTTCATCTTAGCAGTCATTCGTTTCACGGGAAGCCTTGTCCAGTAAGCGGCGGG
CCCATCTGGAGAAATCTCTTCATCAATCGGCTGATGTCCTTCGTTATCGGGGATTCGCGC
ATCTGCAAGGTTGATTGCCACTCGTTCCGGTGTAATCTGCATCCGTCCTCCTGCTTGGCC
GACACAAATTACGATATCTGGCTGATGTTTTTGAATGGCTTGGCGCAGAGTGTCCAGAGC
GGATCTAAAGACGGTTGGAATTTGTTCCGCTGTAATAATGGCTTCTTCTGTCTCGAAGCC
ATTAAGCCGTTTCGCCGCTTCCCATGATGGATTGACGGTTTCTTTGTCAAAAGGGTCAAA
GCCTGTGATCAGCACTTTTTTTCTCATTCTCCCATCTCCTTTTTCTTTTATTCTATTGTT
TATTTATGGGTTTTTCATCAAAATAATGTAAAGGAGTGAATCATAATGGAGCATTTGCCG
GAGCAGTATCGCCAGTTATTCCCAACCTTGCAGACGCATACGATGCTTGCCAGCTGTTCT
CAGAGCGCATTGGCAGAGCCTGTATCAAGGGCGATCCAGGATTATTATGATAGCCTGCTG
TATAAAGGGACGAACTGGAAAGAAGCGATTGAAAAAACAGAGTTTGCGAGAAACGAGTTT
GCAAAGCTGATCGGGGCTGAACCGGATGAAGTGGCGATTGTGCCGTCAGTTTCTGATGCA
CTGGTTTCTGTAGCATCGTCCTTAACTGCATTTGGAAAGAAGCACGTTGTATATACAGAT
ATGGATTTTCCGGCGGTGCCTCATGTTTGGCAGGCACACTCCGATTATACCGTATCCGTC
ATTCCATCAATAGACGGCGTGCTGCCGCTTGAACAATATGAAACGCATATTTCGGATGAA
ACAGTACTGACGTGTGTTCCTCACGTTCATTATCGTGACGGCTATGTTCAGGATATAAAA
GCGATTGCCGAGATTTCTCAGAGAAAGGGCTCTTTATTGTTTGTAGATGCTTATCAATCA
GCCGGGCATATTCCCATTGATGTGAAGGAATGGGGCGTAGATATGCTGGCAGCAGGCACC
CGGAAGTATTTGCTCGGCATACCGGGTGTGGCGTTTCTTTATGTGAGAAAGGAGCTGGCT
GACGCACTGAAGCCGAAAGCATCAGCTTGGTTCGGAAGAGAGAGCGGATTTGATGGGGCT
TATGCAAAAGTCGCGCGCCGTTTTCAAACGGGCACCCCAGCTTTTATCAGCGTATACGCA
GCTGCAGCGGCTTTATCGCTGCTGAATCATATTGGGGTTTCTCATATCAGGGATCATGTG
AAAACGATCTGTGCCGATGCAGTTCAATATGCCGCTGAAAAAGGCCTGCAGCTGGCGGCG
GCACAAGGTGGGATTCAGCCGGGCATGGTTGCGATCCGGGATGAGCGGGCATCGGAAACG
GCGGGGTTGCTGAAGAAGAAAAAAGTGATTTGCGCGCCGCGGGAAAATGTTATCCGTCTC
GCTCCCCATTTTTATAATACGAAGGAGGAAATGCGGCACGCGATTGATGAAATCGCGGCG
AAAACGATCCACAAGTAAACATGAAAAAGCCCCTGAACACTAGTCAGGGGCTTTTCATAT
TAATGATCTACTTTAACGCGTTTCATAAAGAAAGCGCCAATTAAACCGATAATGGCAACA
ATCATTGCAAACACAAATGCGTGCTGTACGCCTGCTGTCAAAGCTTGCGGGATGACTGCC
GGATCGGCAGGGTTTTTAACTGTACTCATATAATCATGCTGGCCTGCAGCCATAATGCTG
ACCGCAACCGCTGTTCCGATAGCGCCGGCCATTTGCTGCAGCGTGTTCATAATGGCGGTG
CCGTCTGGATAAAATTCACGCGGCAGTTGGTTTAAACCGTTTGTCTGTGCAGGCATCATG
ATCATAGAAATCCCGATCATCAAGCAGGTGTGCAGGATGATAATCAGCACAGCTGTTGAA
GTGGTCGTGACATTTGAGAAGAACCATAGTACAACGGTGACAATCACAAATCCCGGAATG
ACAAGCCATTTCGGCCCGTATTTATCGAACAAGCGGCCTGTAACAGGGGACATAAATCCA
TTTAAAATACCGCCCGGCAAGAGAACAAGACCAGATGCAAATGCAGTGAGGACTAAGCCG
CCTTGCAGATACATCGGCAGAAGCAGCATAGATGACAGAATGACCATCATACAAATGAAC
ACCATGATCACACCCAAAATAAACATCGGGTATTTGAACGCACGGAGGTTCATCATAGGC
TGCTTCATTGTCAGCTGGCGGATTGAAAATAAGATAAGGCCGACAACGCCGACAATCAGC
GACACGATAACAGTCGGGCTGGACCATCCCCCGGAGCCTTCACCCGCGTTGCTGAATCCG
AATACAATGCCGCCGAAGCCAATCGTCGACAGGATGATAGACAATACATCGATTTTCGGC
TTTGTCGTTTCAGATACATTTTGCATATATGCGATACCGAAAACAAGCGCCAGCACAAGG
AATGGAAGAGAGATCCAGAAAATCCAGTGCCAGTTGAGATGCTCCAGAACCAATCCTGAG
AAAGTTGGGCCGATGGCGGGCGCGAACATAATGACAAGCCCGATCGTTCCCATTGCGGCA
```

```
CCCCGTTTATGAGGCGGGAAAATCACCAAGATTGTGTTAAACATCAGCGGCAGTAAAAGA
CCGGTTCCAAGTGCCTGAACGATCCTTGCCGCTAATAAAAACGAGAAGCTCGGCGCAAGC
GCCGCAATGAATGTACCTAAAATTGAAAAGATAAGTGACACGGTAAAAAGCTGTCTTGTT
GTGAACCACTGCAACAGCAGTCCTGAAACAGGAACAAGGATACCGAGTACAAGCAGGTAG
CCCGTCGTTAACCATTGGACGGTTGCCGCTGTAATGTTCAATTCCTTCATAAGGTCGGTT
AACGCAATATTCAGCGCTGTTTCACTGAACATGCCGATAAAACCGGCCAACAGCAAGGAA
ATCATAATCGGCATCACTTTGTATTGCTGAGATGCTTTAGCTGTTGTTTCCAAAATCATT
TCCCCTCTCTATCAACTGCATGTAGTATGTCGTTTTTTTTATCTCTTCAGCAGGTCAGGA
ATGCAGCTGGAGATATGAAGGAGCGGCGTACTGTTTTTTGCCGTCAAAGATAAAAGGATG
CCGCCTTCAATCATCGCGTTAACCACAGTGCTGGCTTCTTTTGCACGGCTCTCGCTGCAG
CCAGTCTGCCGCAGTTTTTCCTCATACACAGAGGCCCATTCTTTGTAGGCTTCATGACAG
GCTTCGCGCAACGGTTCGCTTTTCAATGACGTCTCAGCCGCTAGCAAGCCCACAGGCAAG
CCTTCAATGTCTTCCGTACATGAAAACTGGCAGGAGAGCTCCTTCAAAAAGGCTTGAATG
CCTTCCGCTGGATCGGTGCAGGCTTCCATGCAGTCCGCGATTTTCTGACGGATATACTCC
TTCATCTCATTCACGGCTTCGATCGCAAGCTGTTCTTTACCCCCGGGAAAGTGGTAGTAA
AGAGAGCCTTTAGGCGCGCCGCTTTCCTTTATAATCTGGTTCAGCCCCGTGCCGTAATAC
CCTTGCAGCTGAAAAAGCCGGGTAGCTGCCGAAAGGATTTTCTCACGGGAATCTCCATAA
CTCATAACATTCCCACCTTACTGAATTGCAATCAAAAATATAGTGACTGGTCTATTATCT
TGATTCAATCATCAATTGTCAAGAAAAATTCATTGTATGAAAAGACAAAAAAAGAAGGAT
ATGACAACAAAAAATACTGAGAGAAAAGCTGACTGATCTTTTGACTGAATAGATAAAATG
TACAATGATTAATCATCATATGGATGTAAGGAGAGAAATAGATGAAAAAACAACGAATGC
TCGTACTTTTTACCGCACTATTGTTTGTTTTTACCGGATGTTCACATTCTCCTGAAACAA
AAGAATCCCCGAAAGAAAAAGCTCAGACACAAAAAGTCTCTTCGGCTTCTGCCTCTGAAA
AAAAGGATCTGCCAAACATTAGAATTTTAGCGACAGGAGGCACGATAGCTGGTGCCGATC
AATCGAAAACCTCAACAACTGAATATAAAGCAGGTGTTGTCGGCGTTGAATCACTGATCG
AGGCAGTTCCAGAAATGAAGGACATTGCAAACGTCAGCGGCGAGCAGATTGTTAACGTCG
GCAGCACAAATATTGATAATAAAATATTGCTGAAGCTGGCGAAACGCATCAACCACTTGC
TCGCTTCAGATGATGTAGACGGAATCGTCGTGACTCATGGAACAGATACATTGGAGGAAA
CCGCTTATTTTTTGAATCTTACCGTGAAAAGTGATAAACCGGTTGTTATTGTCGGTTCGA
TGAGACCTTCCACAGCCATCAGCGCTGATGGGCCTTCTAACCTGTACAATGCAGTGAAAG
```

De la grande littérature n'est-ce pas ?

Cette liste peut-être téléchargée dans le fichier bsubtilis.285000-290999.txt. Le génome complet (environ quatre millions de nucléotides) est aussi disponible en version compressée.

Une tâche auxiliaire pour exploiter ces données sera d'écrire une petite fonction qui prend en argument le nom d'un fichier au format ci-dessus, et retourne le contenu de ce fichier sous la forme d'une chaîne de nucléotides `'A'`, `'C'`, `'G'` ou `'T'`.

Testez votre algorithme de détection sur ces données conséquentes. Combien trouvez-vous de séquences potentiellement codantes ?

Décodage

Cette deuxième partie est optionnelle mais facile.

On a vu que chaque codon (sauf les trois codons *stop*) code un acide aminé. Mais cette correspondance n'est pas injective : il n'y a qu'une vingtaine d'acides aminés pour 64 codons.

La table suivante est reprise de Wikipédia et un peu modifiée par souci de simplification.

Acide aminé	Codons

Alanine	Ala	A	GCT, GCC, GCA, GCG
Arginine	Arg	R	CGT, CGC, CGA, CGG, AGA, AGG
Asparagine	Asn	N	AAT, AAC
Acide aspartique	Asp	D	GAT, GAC
Cystéine	Cys	C	TGT, TGC
Glutamine	Gln	Q	CAA, CAG
Acide glutamique	Glu	E	GAA, GAG
Glycine	Gly	G	GGT, GGC, GGA, GGG
Histidine	His	H	CAT, CAC
Isoleucine	Ile	I	ATT, ATC, ATA
Leucine	Leu	L	TTA, TTG, CTT, CTC, CTA, CTG
Lysine	Lys	K	AAA, AAG
Méthionine	Met	M	ATG
Phénylalanine	Phe	F	TTT, TTC
Proline	Pro	P	CCT, CCC, CCA, CCG
Sérine	Ser	S	TCT, TCC, TCA, TCG, AGT, AGC
Thréonine	Thr	T	ACT, ACC, ACA, ACG
Tryptophane	Trp	W	TGG
Tyrosine	Tyr	Y	TAT, TAC
Valine	Val	V	GTT, GTC, GTA, GTG
STOP			TAG, TAA, TGA

Écrivez une fonction qui prend en argument un codon et retourne la lettre représentant l'acide aminé produit, ou `None` s'il s'agit d'un codon *stop*.

Déduisez-en une fonction de traduction d'une séquence codante en chaîne d'acides aminés.

Générez toutes les chaînes d'acides aminés correspondant aux séquences potentiellement codantes de l'extrait du génome de *B. subtilis* ci-dessus.

Recherche de concordances avec des protéines connues

Cette troisième partie est absolument optionnelle.

Il s'agit d'identifier parmi les séquences potentiellement codantes celles qui ont déjà été identifiées comme codant des protéines dans des espèces vivantes.

On utilise pour cela la base UniProtKB/Swiss-Prot, qui utilise un format texte à peu près transparent. Voici trois entrées dans cette base :

```
>sp|Q6GZX4|001R_FRG3G Putative transcription factor 001R OS=Frog
```

```
virus 3 (isolate Goorha) GN=FV3-001R PE=4 SV=1
MAFSAEDVLKEYDRRRRMEALLLSLYYPNDRKLLDYKEWSPPRVQVECPKAPVEWNNPPS
EKGLIVGHFSGIKYKGEKAQASEVDVNKMCCWVSKFKDAMRRYQGIQTCKIPGKVLSDLD
AKIKAYNLTVEGVEGFVRYSRVTKQHVAAFLKELRHSKQYENVNLIHYILTDKRVDIQHL
EKDLVKDFKALVESAHRMRQGHMINVKYILYQLLKKHGHGPDGPDILTVKTGSKGVLYDD
SFRKIYTDLGWKFTPL
>sp|Q6GZX3|002L_FRG3G Uncharacterized protein 002L OS=Frog virus 3
(isolate Goorha) GN=FV3-002L PE=4 SV=1
MSIIGATRLQNDKSDTYSAGPCYAGGCSAFTPRGTCGKDWDLGEQTCASGFCTSQPLCAR
IKKTQVCGLRYSSKGKDPLVSAEWDSRGAPYVRCTYDADLIDTQAQVDQFVSMFGESPSL
AERYCMRGVKNTAGELVSRVSSDADPAGGWCRKWYSAHRGPDQDAALGSFCIKNPGAADC
KCINRASDPVYQKVKTLHAYPDQCWYVPCAADVGELKMGTQRDTPTNCPTQVCQIVFNML
DDGSVTMDDVKNTINCDFSKYVPPPPPPKPTPPTPPTPPTPPTPPTPPTPPTPRPVHNRK
VMFFVAGAVLVAILISTVRW
>sp|Q197F8|002R_IIV3 Uncharacterized protein 002R OS=Invertebrate
iridescent virus 3 GN=IIV3-002R PE=4 SV=1
MASNTVSAQGGSNRPVRDFSNIQDVAQFLLFDPIWNEQPGSIVPWKMNREQALAERYPEL
QTSEPSEDYSGPVESLELLPLEIKLDIMQYLSWEQISWCKHPWLWTRWYKDNVVRVSAIT
FEDFQREYAFPEKIQEIHFTDTRAEEIKAILETTPNVTRLVIRRIDDMNYNTHGDLGLDD
LEFLTHLMVEDACGFTDFWAPSLTHLTIKNLDMHPRWFGPVMDGIKSMQSTLKYLYIFET
YGVNKPFVQWCTDNIETFYCTNSYRYENVPRPIYVWVLFQEDEWHGYRVEDNKFHRRYMY
STILHKRDTDWVENNPLKTPAQVEMYKFLLRISQLNRDGTGYESDSDPENEHFDDESFSS
GEEDSSDEDDPTWAPDSDDSDWETETEEEPSVAARILEKGKLTITNLMKSLGFKPKPKKI
QSIDRYFCSLDSNYNSEDEDFEYDSDSEDDDSDSEDDC
```

Le fichier complet est téléchargeable, dans une version compressée au format ZIP, qui pèse déjà 80 Mo : uniprot_sprot.fasta.zip. La version décompressée atteint les 250 Mo environ.

Votre mission, si vous l'acceptez, est de rechercher dans cette base si vos chaînes d'acides aminés s'y trouvent, en fournissant la ligne d'en-tête correspondante. Vous avez carte blanche.

Attention à la manière dont vous gérez ces données : 250 Mo en mémoire ça fait un peu mal. Dix fois 250 Mo, ça fait très mal.

Si ça reste trop facile, vous pouvez vous intéresser à la recherche de protéines *similaires* en suivant les indications du chapitre 3 du document de Pierrick Bouttier.

WikISN: Recherche de séquences codantes dans un génome (dernière édition le 2016-01-21 09:54:19 par LionelVaux)