TD4 : Biologie et Bioinformatique pour Informaticiens – 11 déc 2017

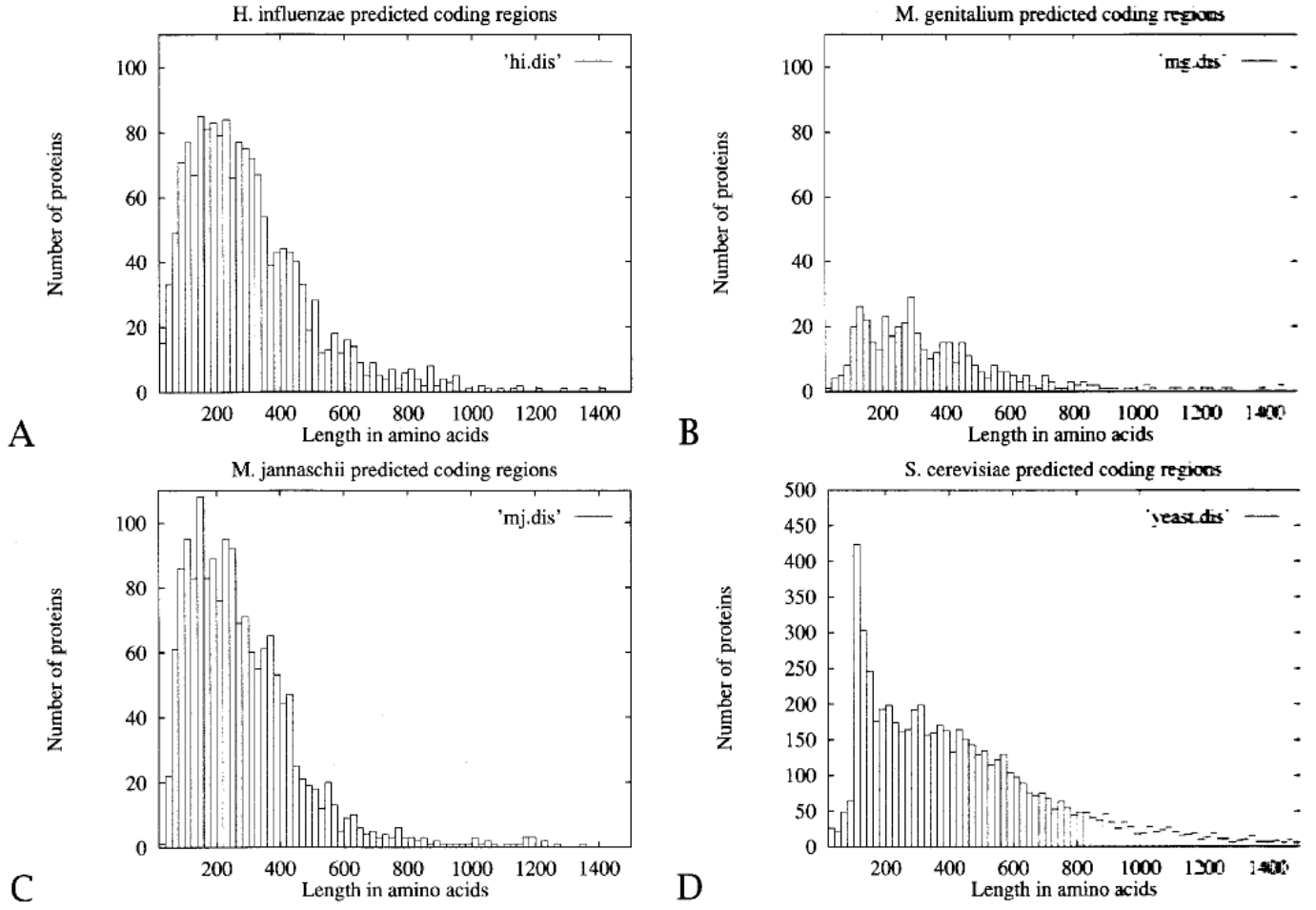

Figure 1.3: Length Distributions for Predicted Protein Coding Regions in Entire Genomes. **A.** *H. influenzae*, among the 1,743 regions, amino acid chains of lengths between 140 and 160 are the most frequent. **B.** *M. genitalium* with 468 regions, and preferred amino acid chains of length between 120 and 140 or 280 and 300. **C.** The archaeon *M. jannaschii* with 1,735 regions; amino acid chains of length between 140 and 160 are the most frequent. **D.** *S. cerevisiae*, among the 6,200 putative protein coding regions, amino acid chains of length between 100 and 120 are the most frequent; this interval is followed by the interval 120 to 140. As described in a 1997 correspondence in *Nature*, the *S. cerevisiae* set clearly contains an overrepresentation (of artifact sequences) in the 100–120 length interval [107].

Soren Brunak et Pierre Baldi dans le livre "Bioinformatics, The Machine Learning Approach" observent la distribution de la taille des gènes pour 4 organismes.

Bacillus_subtilis	NC_000964
Bacillus_cereus_03BB102	NC_012472
Escherichia_coli_K_12_substr__MG1655	NC_000913
Pseudomonas_aeruginosa_LESB58	NC_011770
Lactococcus_lactis_Il1403	NC_002662

1) Distribution des tailles de CDS en AA

A partir des fichiers GBK pour les 5 souches du tableau joint, vous tracerez un histogramme de la taille des CDS en nucléotides. Dans la figure vous indiquerez à partir du fichier GBK, le nom de l'organisme.

Vous utiliserez le code en langage Go (ExtractCDS) – et le package 'gonum/plot'

Quelle modification dans le code du programme fourni avez-vous fait pour permettre l'affichage de l'organisme sur l'histogramme ?

2) Construction d'un décodeur de CDS ultra-simplifié.

A partir du codon MET et des 3 codons STOP, vous parcourrez dans les 6 phases de lecture la séquence nucléotidique de la souche "mystère" et vous tracerez l'histogramme de la taille des CDS en nucléotides.

Vous tiendrez compte du fait que les chromosomes peuvent être circulaires. La souche "mystère" ne sera pas considéré comme circulaire.

Dans la vision ultra-simplifiée, on considère qu'un CDS existe dès lors qu'il débute par un codon MET et fini par un codon STOP. Comme vous travaillez sur les 6 phases de lecture, vous pouvez avoir des recouvrements entre CDS.

Exemple :

GATATAATGAGTTATCAACAAACTAAAAAGTAAAGGAGTAATATGGCATCCCTTAATGAAAATCAAAAA...
CTATATTACTCAATAGTTGTTTGATTTTTCATTTCCTCATTATACCGTAGGGAATTACTTTTAGTTTTT

>Frame +1	DIMSYQQTKK*RSNMASLNENQK
>Frame +2	I**VINKLKSKGVIWHPLMKIK
>Frame +3	YNELSTN*KVKE*YGIP**KSK
>Frame -1	FLIFIKGCHITPLLFSLLITHYI
>Frame -2	F*FSLRDAILLLYFLVC**LII
>Frame -3	FDFH*GMPYYSFTF*FVDNSLY

Sur la séquence Frame+1 : on a un CDS de 8AA et un CDS de 9AA
Sur la séquence Frame+2 : on a 4AA
Sur la séquence Frame+3 : pas de codon Met
Sur la séquence Frame-1 : pas de codon Met
Sur la séquence Frame-2 : pas de codon Met
Sur la séquence Frame-3: on a un CDS de 8AA

1ère position	2nd position U	2nd position C	2nd position A	2nd position G	3ème position
U	PHE	SER	TYR	CYS	U
	PHE	SER	TYR	CYS	C
	LEU	SER	Stop	Stop	A
	LEU	SER	Stop	TRP	G
C	LEU	PRO	HIS	ARG	U
	LEU	PRO	HIS	ARG	C
	LEU	PRO	GLN	ARG	A
	LEU	PRO	GLN	ARG	G
A	ILE	THR	ASN	SER	U
	ILE	THR	ASN	SER	C
	ILE	THR	LYS	ARG	A
	MET	THR	LYS	ARG	G
G	VAL	ALA	ASP	GLY	U
	VAL	ALA	ASP	GLY	C
	VAL	ALA	GLU	GLY	A
	VAL	ALA	GLU	GLY	G

1) **Proposez un modèle d'automate qui gère la détection dans le sens direct et un automate qui gère le sens inverse.**

2) **Tracez la distribution en ne tenant pas compte des gènes de taille inférieure à 500AA**

3) **Comparez vos distributions**

N'oubliez pas de faire une analyse critique (quelques lignes) de vos résultats. Cette analyse critique compte pour moitié des points.