Laboratorio di Elementi di Bioinformatica

Laurea Triennale in Informatica
(codice: E3101Q116)

AA 2016/2017

Esercizio3

Docente del laboratorio: Raffaella Rizzi

Esercizio

Scrivere un programma che prenda in input un file in formato EMBL contenente una sequenza di mRNA (vedere `M10051.txt`), e un file contenente il codice genetico (vedere `genetic-code.txt`) e produca in standard output (in un formato a scelta dello studente):

- la distribuzione di frequenza dei codoni della *coding sequence* (CDS)
- la traduzione della *coding sequence* in sequenza di aminoacidi tramite codice genetico
- la distribuzione di frequenza degli aminoacidi della proteina ottenuta
- se la traduzione ottenuta è uguale alla traduzione riportata nel file di input

Esercizio

Scrivere un programma che prenda in input un file in formato EMBL contenente una sequenza di mRNA (vedere `M10051.txt`), e un file contenente il codice genetico (vedere `genetic-code.txt`) e produca in standard output (in un formato a scelta dello studente):

- la distribuzi*ence* (CDS)
- la traduzione della *coding sequence* in sequenza di aminoacidi tramite codice genetico
- la distribuzione di frequenza degli aminoacidi della proteina ottenuta
- se la traduzione ottenuta è uguale alla traduzione riportata nel file di input

Per un esempio di output vedere il file `output.txt`

mRNA → CDS → proteina

L'mRNA è il prodotto dell'espressione di un gene e la *coding sequence* è la sottostringa di mRNA che viene tradotta in proteina.

mRNA → CDS → proteina

L'mRNA è il prodotto dell'espressione di un gene e la *coding sequence* è la sottostringa di mRNA che viene tradotta in proteina.

Una proteina è una sequenza di aminoacidi e risulta rappresentata da una sequenza di simboli appartenenti a un alfabeto di dimensione 20.

mRNA → CDS → proteina

L'mRNA è il prodotto dell'espressione di un gene e la *coding sequence* è la sottostringa di mRNA che viene tradotta in proteina.

Una proteina è una sequenza di aminoacidi e risulta rappresentata da una sequenza di simboli appartenenti a un alfabeto di dimensione 20.

CDS →	proteina
{a,c,g,t}	{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}

mRNA → CDS → proteina

L'mRNA è il prodotto dell'espressione di un gene e la *coding sequence* è la sottostringa di mRNA che viene tradotta in proteina.

Una proteina è una sequenza di aminoacidi e risulta rappresentata da una sequenza di simboli appartenenti a un alfabeto di dimensione 20.

mRNA **gtaagcatgccaagcgattaggttg**

La CDS ha le seguenti proprietà:

- la sua lunghezza è un multiplo di tre
- inizia con **atg** e finisce con **tag** | **taa** | **tga**

mRNA → CDS → proteina

mRNA **gtaagcatgccaagcgattaggttg**

CDS **atgccaagcgattag**

mRNA → CDS → proteina

mRNA **gtaagcatgccaagcgattaggttg**

CDS **atgccaagcgattag**

La CDS deve essere pensata come una sequenza di triplette (chiamate codoni)

atg cca agc gat tag

mRNA → CDS → proteina

mRNA **gtaagcatgccaagcgattaggttg**

CDS **atgccaagcgattag**

La CDS deve essere pensata come una sequenza di triplette (chiamate codoni)

atg cca agc gat tag

Ogni codone, attraverso il codice genetico, viene tradotto in un aminoacido, ottenendo quindi una proteina

mRNA → CDS → proteina

Il codice genetico è una tabella che mappa ad un aminoacido (tra i venti esistenti in natura) ognuna delle 4^3 possibili triplette (codoni) che si possono formare con l'alfabeto {a,c,g,t}

Prima base	Seconda base								Terza base
	T		C		A		G		
T	TTT	(Phe/F) Fenilalanina	TCT	(Ser/S) Serina	TAT	(Tyr/Y) Tirosina	TGT	(Cys/C) Cisteina	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucina	TCA		TAA	Stop (*Ocra*)	TGA	Stop (*Opale*)	A
	TTG		TCG		TAG	Stop (*Ambra*)	TGG	(Trp/W) Triptofano	G
C	CTT		CCT	(Pro/P) Prolina	CAT	(His/H) Istidina	CGT	(Arg/R) Arginina	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutammina	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucina	ACT	(Thr/T) Treonina	AAT	(Asn/N) Asparagina	AGT	(Ser/S) Serina	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lisina	AGA	(Arg/R) Arginina	A
	ATG	(Met/M) Metionina	ACG		AAG		AGG		G
G	GTT	(Val/V) Valina	GCT	(Ala/A) Alanina	GAT	(Asp/D) Acido aspartico	GGT	(Gly/G) Glicina	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Acido glutammico	GGA		A
	GTG		GCG		GAG		GGG		G

mRNA → CDS → proteina

mRNA **gtaagcatgccaagcgattaggttg**

CDS **atgccaagcgattag**

La CDS deve essere pensata come una sequenza di triplette (chiamate codoni)

atg cca agc gat tag

Ogni codone, attraverso il codice genetico, viene tradotto in un aminoacido, ottenendo quindi una proteina

MPSD[stop]

Il formato EMBL

Il *record* “FT” seguito da spazi e dalla stringa “/translation=“ è il primo di una serie di record “FT” contenenti la traduzione della CDS.

```
FT                   /protein_id="AAA59174.1"
FT                   /translation="MGTGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLT
FT                   RLHELENCSVIEGHLQILLMFKTRPEDFRDLSFPKLIMITDYLLLFRVYGLESLKDLFP
FT                   NLTVIRGSRLFFNYALVIFEMVHLKELGLYNLMNITRGSVRIEKNNELCYLATIDWSRI
FT                   LDSVEDNHIVLNKDDNEECGDICPGTAKGKTNCPATVINGQFVERCWTHSHCQKVCPTI
FT                   CKSHGCTAEGLCCHSECLGNCSQPDDPTKCVACRNFYLDGRCVETCPPPYYHFQDWRCV
FT                   NFSFCQDLHHKCKNSRRQGCHQYVIHNNKCIPECPSGYTMNSSNLLCTPCLGPCPKVCH
FT                   LLEGEKTIDSVTSAQELRGCTVINGSLIINIRGGNNLAAELEANLGLIEEISGYLKIRR
FT                   SYALVSLSFFRKLRLIRGETLEIGNYSFYALDNQNLRQLWDWSKHNLTTTQGKLFFHYN
FT                   PKLCLSEIHKMEEVSGTKGRQERNDIALKTNGDKASCENELLKFSYIRTSFDKILLRWE
FT                   PYWPPDFRDLLGFMLFYKEAPYQNVTEFDGQDACGSNSWTVVDIDPPLRSNDPKSQNHP
FT                   GWLMRGLKPWTQYAIFVKTLVTFSDERRTYGAKSDIIYVQTDATNPSVPLDPISVSNSS
FT                   SQIILKWKPPSDPNGNITHYLVFWERQAEDSELFELDYCLKGLKLPSRTWSPPFESEDS
FT                   QKHNQSEYEDSAGECCSCPKTDSQILKELEESSFRKTFEDYLHNVVFVPRKTSSGTGAE
FT                   DPRPSRKRRSLGDVGNVTVAVPTVAAFPNTSSTSVPTSPEEHRPFEKVVNKESLVISGL
FT                   RHFTGYRIELQACNQDTPEERCSVAAYVSARTMPEAKADDIVGPVTHEIFENNVVHLMW
FT                   QEPKEPNGLIVLYEVSYRRYGDEELHLCVSRKHFALERGCRLRGLSPGNYSVRIRATSL
FT                   AGNGSWTEPTYFYVTDYLDVPSNIAKIIIGPLIFVFLFSVVIGSIYLFLRKRQPDGPLG
FT                   PLYASSNPEYLSASDVFPCSVYVPDEWEVSREKITLLRELGQGSFGMVYEGNARDIIKG
FT                   EAETRVAVKTVNESASLRERIEFLNEASVMKGFTCHHVVRLLGVVSKGQPTLVVMELMA
FT                   HGDLKSYLRSLRPEAENNPGRPPPTLQEMIQMAAEIADGMAYLNAKKFVHRDLAARNCM
FT                   VAHDFTVKIGDFGMTRDIYETDYYRKGGKGLLPVRWMAPESLKDGVFTTSSDMWSFGVV
FT                   LWEITSLAEQPYQGLSNEQVLKFVMDGGYLDQPDNCPERVTDLMRMCWQFNPKMRPTFL
FT                   EIVNLLKDDLHPSFPEVSFFHSEENKAPESEELEMEFEDMENVPLDRSSHCQREEAGGR
FT                   DGGSSLGFKRSYEEHIPYTHMNGGKKNGRILTLPRSNPS"
FT   mat_peptide     220..2424
FT                   /gene="INSR"
```

Il formato EMBL

Il *record* “FT” seguito da spazi e dalla stringa “/translation=“ è il primo di una serie di record “FT” contenenti la traduzione della CDS.

```
FT                   /protein_id="AAA59174.1"
FT                   /translation="MGTGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLT
FT                   RLHELENCSVIEGHLQILLMFKTRPEDFRDLSFPKLIMITDYLLLFRVYGLESLKDLFP
FT                   NLTVIRGSRLFFNYALVIFEMVHLKELGLYNLMNITRGSVRIEKNNELCYLATIDWSRI
FT                   LDSVEDNHIVLNKDDNEECGDICPGTAKGKTNCPATVINGQFVERCWTHSHCQKVCPTI
```

Uno dei pattern che riconoscono i record che contengono la sequenza della proteina è:

`/^FT\s+(\/translation=“)?([ACDEFGHIKLMNPQRSTVWY]+)”?$/`

`$2` → chunk di proteina

```
FT                   RHFTGYRIELQACNQDTPEERCSVAAYVSARTMPEAKADDIVGPVTHEIFENNVVHLMW
FT                   QEPKEPNGLIVLYEVSYRRYGDEELHLCVSRKHFALERGCRLRGLSPGNYSVRIRATSL
FT                   AGNGSWTEPTYFYVTDYLDVPSNIAKIIIGPLIFVFLFSVVIGSIYLFLRKRQPDGPLG
FT                   PLYASSNPEYLSASDVFPCSVYVPDEWEVSREKITLLRELGQGSFGMVYEGNARDIIKG
FT                   EAETRVAVKTVNESASLRERIEFLNEASVMKGFTCHHVVRLLGVVSKGQPTLVVMELMA
FT                   HGDLKSYLRSLRPEAENNPGRPPPTLQEMIQMAAEIADGMAYLNAKKFVHRDLAARNCM
FT                   VAHDFTVKIGDFGMTRDIYETDYYRKGGKGLLPVRWMAPESLKDGVFTTSSDMWSFGVV
FT                   LWEITSLAEQPYQGLSNEQVLKFVMDGGYLDQPDNCPERVTDLMRMCWQFNPKMRPTFL
FT                   EIVNLLKDDLHPSFPEVSFFHSEENKAPESEELEMEFEDMENVPLDRSSHCQREEAGGR
FT                   DGGSSLGFKRSYEEHIPYTHMNGGKKNGRILTLPRSNPS"
FT   mat_peptide     220..2424
FT                   /gene="INSR"
```