Laboratorio di Elementi di Bioinformatica

Laurea Triennale in Informatica
(codice: E3101Q116)

AA 2016/2017

Esercizio2

Docente del laboratorio: Raffaella Rizzi

Esercizio

Scrivere un programma che prenda in input un file in formato EMBL (vedere il file `M10051.txt`) e produca in standard output in formato FASTA la sequenza nucleotidica e, nel caso in cui l'*entry* si riferisca a un mRNA, la sequenza della coding sequence (CDS).

Nell'header della sequenza nucleotidica devono comparire (in un formato a scelta dello studente):

- l'accession number (AC) e la lunghezza della sequenza
- la descrizione della sequenza
- l'organismo a cui appartiene la sequenza

Nell'header della coding sequence devono comparire (in un formato a scelta dello studente):

- lo start e l'end rispetto alla sequenza nucleotidica dell'entry
- il nome del gene a cui appartiene l'mRNA
- la presenza nella CDS dello start codon `atg`
- la presenza nella CDS dello stop codon {`tag`, `taa`, `tga`}

Esercizio

Scrivere un programma che prenda in input un file in formato EMBL (vedere il file `M10051.txt`) e produca in standard output in formato FASTA la sequenza nucleotidica e, nel caso in cui l'*entry* si riferisca a un mRNA, la sequenza della coding sequence (CDS).

Nell'header de
scelta dello st

Il nome del file di input deve essere specificato da linea di comando

- l'accession number (AC) e la lunghezza della sequenza
- la descrizione della sequenza
- l'organismo a cui appartiene la sequenza

Nell'header della coding sequence devono comparire (in un formato a scelta dello studente):

- lo start e l'end rispetto alla sequenza nucleotidica dell'entry
- il nome del gene a cui appartiene l'mRNA
- la presenza nella CDS dello start codon `atg`
- la presenza nella CDS dello stop codon {`tag`, `taa`, `tga`}

Esercizio

Scrivere un programma che prenda in input un file in formato EMBL (vedere il file `M10051.txt`) e produca in standard output in formato FASTA la sequenza nucleotidica e, nel caso in cui l'*entry* si riferisca a un mRNA, la sequenza della coding sequence (CDS).

Nell'header de
scelta dello st

Le sequenze devono essere spezzate in righe di 80 caratteri. Le sequenze devono essere in lettere maiuscole.

- l'accession number (AC) e la lunghezza della sequenza
- la descrizione della sequenza
- l'organismo a cui appartiene la sequenza

Nell'header della coding sequence devono comparire (in un formato a scelta dello studente):

- lo start e l'end rispetto alla sequenza nucleotidica dell'entry
- il nome del gene a cui appartiene l'mRNA
- la presenza nella CDS dello start codon `atg`
- la presenza nella CDS dello stop codon {`tag`, `taa`, `tga`}

Esercizio

Scrivere un programma che prenda in input un file in formato EMBL (vedere il file `M10051.txt`) e produca in standard output in formato FASTA la sequenza nucleotidica e, nel caso in cui l'*entry* si riferisca a un mRNA, la sequenza della coding sequence (CDS).

Nell'header de
scelta dello studente):

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

- l'accession number (AC) e la lunghezza della sequenza
- la descrizione della sequenza
- l'organismo a cui appartiene la sequenza

Nell'header della coding sequence devono comparire (in un formato a scelta dello studente):

- lo start e l'end rispetto alla sequenza nucleotidica dell'entry
- il nome del gene a cui appartiene l'mRNA
- la presenza nella CDS dello start codon `atg`
- la presenza nella CDS dello stop codon {`tag`, `taa`, `tga`}

Il formato EMBL

EMBL è un formato di puro testo composto da *record* identificati da un codice a due caratteri maiuscoli nelle prime due posizioni (e seguiti da almeno tre spazi). In particolare:

- accession number (AC) e lunghezza della sequenza
  - Il record "ID" contiene una serie di campi separati da punto e virgola, di cui il primo è l'accession number AC e l'ultimo riporta la lunghezza della sequenza:

    ID **M10051**; SV 1; linear; mRNA; STD; HUM; **4723 BP**.
- descrizione della sequenza
  - Il record "DE" contiene la descrizione della sequenza

    DE **Human insulin receptor mRNA, complete cds**.
- organismo a cui appartiene la sequenza
  - Il record "OS" contiene l'organismo

    OS **Homo sapiens (human)**

Il formato EMBL

EMBL è un formato di puro testo composto da *record* identificati da un codice a due caratteri maiuscoli nelle prime due posizioni (e seguiti da almeno tre spazi). In particolare:

- start ed end della (eventuale) sequenza codificante (CDS)
  - Il record “FT”, seguito da spazi e dalla stringa “CDS”, contiene lo start e l’end della CDS sulla sequenza:

    FT CDS **139..4287**
- nome del gene che esprime la CDS
  - Il record “FT”, seguito da spazi e dalla stringa “/gene=”, contiene il nome del gene

    FT /gene=**"INSR**”
- sequenza nucleotidica
  - La sequenza nucleotidica è contenuta nella parte compresa tra il record che inizia con “SQ” e il record che inizia con // (fine del file).

Il formato EMBL

EMBL è un formato di puro testo composto da *record* identificati da un codice a due caratteri maiuscoli nelle prime due posizioni (e seguiti da almeno tre spazi). In particolare:

- start ed end della (eventuale) sequenza codificante (CDS)
  - Il record "FT", seguito da spazi e dalla stringa "CDS", contiene lo start e l'end della CDS sulla sequenza:

    FT CDS **139..4287**
- nome del gene che esprime la CDS
  - Il record "FT", seguito da spazi e dalla stringa "/gene=", contiene il nome del gene
- se

Tutti i record contenenti la sequenza nucleotidica iniziano con degli spazi e non con un codice a due lettere maiuscole

  - La sequenza nucleotidica è contenuta nella parte compresa tra il record che inizia con "SQ" e il record che inizia con // (fine del file).

Il formato EMBL

```
ID   M10051; SV 1; linear; mRNA; STD; HUM; 4723 BP.
XX
AC
XX
DT
DT   14-NOV-2006 (Rel. 89, Last updated, Version 7)
XX
DE   Human insulin receptor mRNA, complete cds.
XX
KW   insulin receptor; tyrosine kinase.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-4723
RX   DOI; 10.1016/0092-8674(85)90334-4.
RX   PUBMED; 2859121.
RA   Ebina Y., Ellis L., Jarnagin K., Edery M., Graf L., Clauser E., Ou J.-H.,
RA   Masiarz F., Kan Y.W., Goldfine I.D., Roth R.A., Rutter W.J.;
RT   "The human insulin receptor cDNA: the structural basis for
RT   hormone-activated transmembrane signalling";
RL   Cell 40(4):747-758(1985).
XX
```

Accession number → M10051
Lunghezza → 4723

Il formato EMBL

```
ID   M10051; SV 1; linear; mRNA; STD; HUM; 4723 BP.
XX
AC
XX
DT
DT   14-NOV-2006 (Rel. 89, Last updated, Version 7)
XX
DE   Human insuli
XX
KW   insulin rece
XX
OS   Homo sapiens
OC   Eukaryota; M
OC   Eutheria; Eu
OC   Homo.
XX
RN   [1]
RP   1-4723
RX   DOI; 10.1016/0092-8674(85)90334-4.
RX   PUBMED; 2859121.
RA   Ebina Y., Ellis L., Jarnagin K., Edery M., Graf L., Clauser E., Ou J.-H.,
RA   Masiarz F., Kan Y.W., Goldfine I.D., Roth R.A., Rutter W.J.;
RT   "The human insulin receptor cDNA: the structural basis for
RT   hormone-activated transmembrane signalling";
RL   Cell 40(4):747-758(1985).
XX
```

Accession number → M10051
Lunghezza → 4723

Uno dei pattern che riconoscono il record “ID” e ne estraggono Accession Number e lunghezza è:

```
/^ID\s+(\w+).+?(\d+)\s*(BP|bp)/
```

$1 → AC
$2 → lunghezza

Il formato EMBL

```
ID   M10051; SV 1; linear; mRNA; STD; HUM; 4723 BP.
XX
AC   M10051;
XX
DT   02-JUL-1986 (Rel. 09, Created)
DT   14-NOV-2006 (Rel. 89, Last updated, Version 7)
XX
DE   Human insulin receptor mRNA, complete cds.
XX
KW
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-4723
RX   DOI; 10.1016/0092-8674(85)90334-4.
RX   PUBMED; 2859121.
RA   Ebina Y., Ellis L., Jarnagin K., Edery M., Graf L., Clauser E., Ou J.-H.,
RA   Masiarz F., Kan Y.W., Goldfine I.D., Roth R.A., Rutter W.J.;
RT   "The human insulin receptor cDNA: the structural basis for
RT   hormone-activated transmembrane signalling";
RL   Cell 40(4):747-758(1985).
XX
```

Descrizione → “Human insulin receptor mRNA, complete cds”

Il formato EMBL

```
ID   M10051; SV 1; linear; mRNA; STD; HUM; 4723 BP.
XX
AC   M10051;
XX
DT   02-JUL-1986 (Rel. 09, Created)
DT   14-NOV-2006 (Rel. 89, Last updated, Version 7)
XX
DE   Human insulin receptor mRNA, complete cds.
XX
KW
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria;
OC   Homo.
XX
RN   [1]
RP   1-4723
RX   DOI; 10.
RX   PUBMED;
RA   Ebina Y.,
RA   Masiarz F.,
RT   "The human insulin receptor cDNA: the structural basis for
RT   hormone-activated transmembrane signalling";
RL   Cell 40(4):747-758(1985).
XX
```

Descrizione → “Human insulin receptor mRNA, complete cds”

Il pattern che riconosce il record “DE” e ne estrae la descrizione è:

`/^DE\s+(.+)\./`

`$1` → descrizione

Il formato EMBL

```
ID   M10051; SV 1; linear; mRNA; STD; HUM; 4723 BP.
XX
AC   M10051;
XX
DT   02-JUL-1986 (Rel. 09, Created)
DT   14-NOV-2006 (Rel. 89, Last updated, Version 7)
XX
DE   Human insulin receptor mRNA, complete cds.
XX
KW   insulin receptor; tyrosine kinase.
XX
OS   Homo sapiens (human)
OC                                                  ta; Euteleostomi; Mammalia;
OC                                                  i; Catarrhini; Hominidae;
OC
XX
RN   [1]
RP   1-4723
RX   DOI; 10.1016/0092-8674(85)90334-4.
RX   PUBMED; 2859121.
RA   Ebina Y., Ellis L., Jarnagin K., Edery M., Graf L., Clauser E., Ou J.-H.,
RA   Masiarz F., Kan Y.W., Goldfine I.D., Roth R.A., Rutter W.J.;
RT   "The human insulin receptor cDNA: the structural basis for
RT   hormone-activated transmembrane signalling";
RL   Cell 40(4):747-758(1985).
XX
```

Organismo → “Homo sapiens (human)”

Il formato EMBL

```
ID   M10051; SV 1; linear; mRNA; STD; HUM; 4723 BP.
XX
AC   M10051;
XX
DT   02-JUL-1986 (Rel. 09, Created)
DT   14-NOV-2006 (Rel. 89, Last updated, Version 7)
XX
DE   Human insulin receptor mRNA, complete cds.
XX
KW   insulin receptor; tyrosine kinase.
XX
OS   Homo sapiens (human)
OC                                                  ta; Euteleostomi; Mammalia;
OC                                                  i; Catarrhini; Hominidae;
OC
XX
RN   [1]
RP   1-4723
RX   DOI; 10.1016
RX   PUBMED; 2859
RA   Ebina Y., El
RA   Masiarz F.,
RT   "The human in
RT   hormone-activated transmembrane signalling";
RL   Cell 40(4):747-758(1985).
XX
```

Organismo → “Homo sapiens (human)”

Il pattern che riconosce il record “OS” e ne estrae l’organismo è:

`/^OS\s+(.+)/`

`$1` → organismo

Il formato EMBL

```
CC   cleavage product produced upon binding of insulin.  [1] suggests
CC   that translation may begin at the 'atg' start codon at positions
CC   79-81 with protein cleavage occurring after position 120 to yield
CC   the signal peptide.  [1] gives illustrations of the various domains
CC   present in the protein.  A draft entry and sequence for [1] in
CC   computer-readable form were kindly provided by K. Jarnagin
CC   (30-JUL-1985).
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..4723
FT                   /organism="Homo sapiens"
FT                   /map="19p13.3-p13.2"
FT                   /mol_type="mRNA"
FT                   /db_xref="taxon:9606"
FT   sig_peptide     137..219
FT                   /note="insulin receptor signal peptide"
FT   CDS             139..4287
FT
FT
FT
FT                   /db_xref="H-InvDB:HIT000194074.15"
```

CDS → sottostringa della sequenza nucleotidica che va dal carattere in posizione 139 al carattere in posizione 4287

Il formato EMBL

```
CC   cleavage product produced upon binding of insulin.  [1] suggests
CC   that translation may begin at the 'atg' start codon at positions
CC   79-81 with protein cleavage occurring after position 120 to yield
CC   the signal peptide.  [1] gives illustrations of the various domains
CC   present in the protein.  A draft entry and sequence for [1] in
CC   computer-readable form were kindly provided by K. Jarnagin
CC   (30-JUL-1985).
XX
FH   Key
FH
FT   source
FT
FT
FT
FT
FT   sig_peptide
FT                   /note="insulin receptor signal peptide"
FT   CDS             139..4287
FT
FT
FT
FT
FT                   /db_xref="H-InvDB:HIT000194074.15"
```

Uno dei pattern che riconoscono il record “FT” contenente start ed end della CDS e li estraggono è:

`/^FT\s+CDS\s+(\d+)\.\.(\d+)/`

$1 → start

$2 → end

CDS → sottostringa della sequenza nucleotidica che va dal carattere in posizione 139 al carattere in posizione 4287

Il formato EMBL

```
CC   cleavage product produced upon binding of insulin.  [1] suggests
CC   that translation may begin at the 'atg' start codon at positions
CC   79-81 with protein cleavage occurring after position 120 to yield
CC   the signal peptide.  [1] gives illustrations of the various domains
CC   present in the protein.  A draft entry and sequence for [1] in
CC   computer-readable form were kindly provided by K. Jarnagin
CC   (30-JUL-1985).
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..4723
FT                   /organism="Homo sapiens"
FT                   /map="19p13.3-p13.2"
FT                   /mol_type="mRNA"
FT                   /db_xref="taxon:9606"
FT   sig_peptide     137..219
FT                   /note="insulin receptor signal peptide"
FT   CDS             139..4287
FT                   /codon_start=1
FT                   /gene="INSR"
FT
FT
FT                   /db_xref="H-InvDB:HIT000194074.15"
```

Gene → INSR

Il formato EMBL

```
CC   cleavage product produced upon binding of insulin.  [1] suggests
CC   that translation may begin at the 'atg' start codon at positions
CC   79-81 with protein cleavage occurring after position 120 to yield
CC   the signal peptide.  [1] gives illustrations of the various domains
CC   present in the protein.  A draft entry and sequence for [1] in
CC   computer-readable form were kindly provided by K. Jarnagin
CC   (30-JUL-1985).
XX
FH   Key
FH
FT   source
FT
FT
FT
FT
FT   sig_peptide     137..219
FT                   /note="insulin receptor signal peptide"
FT   CDS             139..4287
FT                   /codon_start=1
FT                   /gene="INSR"
FT
FT
FT                   /db_xref="H-InvDB:HIT000194074.15"
```

Uno dei pattern che riconoscono il record “FT” contenente il nome del gene e lo estraggono è:

`/^FT\s+\/gene=.(\w+)./`

`$1` → gene

Gene → INSR

Il formato EMBL

```
SQ   Sequence 4723 BP; 1068 A; 1298 C; 1311 G; 1046 T; 0 other;
     gggggctgc gcggccgggt cggtgcgcac acgagaagga cgcgcggccc ccagcgctct        60
     tgggggccgc ctcggagcat gacccccgcg ggccagcgcc gcgcgcctga tccgaggaga       120
     ccccgcgctc ccgcagccat gggcaccggg ggccggcggg gggcggcggc cgcgccgctg       180
     ctggtggcgg tggccgcgct gctactgggc gccgcgggcc acctgtaccc cggagaggtg       240
     tgtcccggca tggatatccg gaacaacctc actaggttgc atgagctgga gaattgctct       300
     gtcatcgaag gacacttgca gatactcttg atgttcaaaa cgaggcccga agatttccga       360
     gacctcagtt tccccaaact catcatgatc actgattact tgctgctctt ccgggtctat       420
     gggctcgaga gcctgaagga cctgttcccc aacctcacgg tcatccgggg atcacgactg       480
     ttctttaact acgcgctggt catcttcgag atggttcacc tcaaggaact cggcctctac       540
     aacctgatga acatcacccg gggttctgtc cgcatcgaga agaacaatga gctctgttac       600
     ttggccacta tcgactggtc ccgtatcctg gattccgtgg aggataatca catcgtgttg       660
     aacaaagatg acaacgagga gtgtggagac atctgtccgg gtaccgcgaa gggcaagacc       720
     aactgccccg ccaccgtcat caacgggcag tttgtcgaac gatgttggac tcatagtcac       780
     tgccagaaag tttgcccgac catctgtaag tcacacggct gcaccgccga aggcctctgt       840
     tgccacagcg agtgcctggg caactgttct cagcccgacg accccaccaa gtgcgtggcc       900
     tgccgcaact tctacctgga cggcaggtgt gtggagacct gcccgccccc gtactaccac       960
     ttccaggact ggcgctgtgt gaacttcagc ttctgccagg acctgcacca caaatgcaag      1020
     aactcgcgga ggcagggctg ccaccaatac gtcattcaca acaacaagtg catccctgag      1080
     aacaatctgg cagctgagct agaagccaac ctcggcctca ttgaagaaat ttcagggtat      1320
     ctaaaaatcc gccgatccta cgctctggtg tcactttcct tcttccggaa gttacgtctg      1380
     attcgaggag agaccttgga aattgggaac tactccttct atgccttgga caaccagaac      1440
     ctaaggcagc tctgggactg gagcaaacac aacctcacca ccactcaggg gaaactcttc      1500
```

Sequenza nucleotidica →

ggggggctgcgcggccgggtcggtgcgcacacgagaaggacgcgcggccc…

Il formato EMBL

```
SQ   Sequence 4723 BP; 1068 A; 1298 C; 1311 G; 1046 T; 0 other;
     gggggctgc gcggccgggt cggtgcgcac acgagaagga cgcgcggccc ccagcgctct        60
     tgggggccgc ctcggagcat gacccccgcg ggccagcgcc gcgcgcctga tccgaggaga       120
     ccccgcgctc ccgcagccat gggcaccggg ggccggcggg gggcggcggc cgcgccgctg       180
     ctggtggcgg tggccgcgct gctactgggc gccgcgggcc acctgtaccc cggagaggtg       240
     tgtcccggca tggatatccg gaacaacctc actaggttgc atgagctgga gaattgctct       300
     gtcatcgaag gacacttgca gatactcttg atgttcaaaa cgaggcccga agatttccga       360
     gacctcagtt tccccaaact catcatgatc actgattact tgctgctctt ccgggtctat       420
     gggctcgaga gcctgaagga cctgttcccc aacctcacgg tcatccgggg atcacgactg       480
     ttctttaact a
     aacctgatga a
     ttggccacta t
     aacaaagatg a
     aactgccccg c
     tgccagaaag t
     tgccacagcg agtgcctggg caactgttct cagcccgacg accccaccaa gtgcgtggcc       900
     tgccgcaact tctacctgga cggcaggtgt gtggagacct gcccgccccc gtactaccac       960
     ttccaggact ggcgctgtgt gaacttcagc ttctgccagg acctgcacca caaatgcaag      1020
     aactcgcgga ggcagggctg ccaccaatac gtcattcaca acaacaagtg catccctgag      1080
     aacaatctgg cagctgagct agaagccaac ctcggcctca ttgaagaaat ttcagggtat      1320
     ctaaaaatcc gccgatccta cgctctggtg tcactttcct tcttccggaa gttacgtctg      1380
     attcgaggag agaccttgga aattgggaac tactccttct atgccttgga caaccagaac      1440
     ctaaggcagc tctgggactg gagcaaacac aacctcacca ccactcaggg gaaactcttc      1500
```

Uno dei pattern che riconoscono il record che contengono la sequenza nucleotidica è:

```
/^\s+/
```

Sequenza nucleotidica →

```
gggggctgcgcggccgggtcggtgcgcacacgagaaggacgcgcggccc…
```

Suggerimenti

- il metodo **`start_with?`** della classe `String` prende come argomenti un numero qualsiasi di stringhe e restituisce `true` se l'oggetto invocante inizia con una delle stringhe passate come argomento (altrimenti restituisce `false`)
- il metodo **`end_with?`** della classe `String` prende come argomenti un numero qualsiasi di stringhe e restituisce `true` se l'oggetto invocante finisce con una delle stringhe passate come argomento (altrimenti restituisce `false`)
- il metodo **`split(`*`sep`*`)`** della classe `String` divide la stringa usando *`sep`* come separatore e restituisce un array contenente le singole parti
- ***`string_name`*`[`*`start, length`*`]`** permette di ottenere la sottostringa di *`length`* caratteri di *`string_name`* che inizia in *`start`*

Suggerimenti

- il metodo **`chomp!`** della classe `String` rimuove il carattere di *newline* eventualmente presente alla fine della stringa invocante