

Trabajo Práctico 1

16.50 - Introducción a la Bioinformática

Integrantes:

Cortés Rodríguez, Kevin

Lobo, Daniel

Vega, Magdalena

Instituto Tecnológico de Buenos Aires
1er. Cuatrimestre 2018

Fecha límite de entrega: Lunes, 28 de Mayo

Ejercicio 1: Procesamiento de secuencias

- Ejercicio realizado en dos lenguajes: BioRuby y BioPerl
- Objetivo: Realizar un script, en los lenguajes ya mencionados, que tenga por input un archivo de GenBank y traducirlo a un formato “FASTA” como una secuencia de aminoácidos.

GenBank es una base de datos pública que contiene una extensa colección de secuencias de nucleótidos obtenidas a partir de más de 300.000 especies. Además de la secuencia, incluye información bibliográfica, anotaciones funcionales y, si se trata de una secuencia codificante, su traducción conceptual a proteína.

Cada registro de GenBank contiene 4 secciones:

- Encabezamiento:** Contiene información general sobre el registro, distribuida en varias líneas de información. El nombre de cada línea, así como la información que contiene se resume:
 - i. Locus
 - ii. Definition
 - iii. Accession
 - iv. Version
 - v. Keywords
 - vi. Source
 - vii. Organism
- Referencias bibliográficas:** Cada registro contiene por lo menos una referencia bibliográfica que incluye el nombre de los autores, el título del artículo, la revista donde se ha publicado y el identificador de PubMed (PMID). Cuando hay más de una aparecen numeradas y se muestran por orden cronológico, comenzando por las más antiguas. La última referencia contiene información sobre los autores que han enviado la secuencia a GenBank y la fecha del envío. Contiene las siguientes secciones:
 - i. Reference
 - ii. Comments
- Tabla de características**
 - i. Keywords
- Secuencias:** En este apartado se incluye la secuencia completa. Cada línea tiene 60 nucleótidos dispuestos en 6 bloques de 10.

FASTA, a diferencia, es un formato basado en texto, utilizado para representar secuencias de ácidos nucleicos, péptido, o aminoácidos. El formato también permite incluir nombres de secuencias y comentarios que preceden a las secuencias en sí.

Como decía el enunciado, se requería que la salida solicitada por el programa sea como una secuencia de aminoácidos. Los códigos de aminoácidos aceptados son:

A ALA alanine	P PRO proline
B ASX aspartate or asparagine	Q GLN glutamine

C CYS cystine	R ARG arginine
D ASP aspartate	S SER serine
E GLU glutamate	T THR threonine
F PHE phenylalanine	U selenocysteine
G GLY glycine	V VAL valine
H HIS histidine	W TRP tryptophan
I ILE isoleucine	Y TYR tyrosine
K LYS lysine	Z GLX glutamate or glutamine
L LEU leucine	X any
M MET methionine	* translation stop
N ASN asparagine	- gap of indeterminate length

- Código:

```
#!/usr/bin/env ruby

require 'bio'

if ARGV.length != 2
  puts 'Parametros: <NOMBRE.gb> <SALIDA>'
  exit
end

entries = Bio::GenBank.open(ARGV[0])

File.open(ARGV[1], 'w') do |f|
  entries.each_entry do |entry|
    definition = "#{entry.definition}"
    val = entry.seq.translate
    f.puts val.to_fasta(definition)
  end
end
```

Primero se importa la librería “bio”, para hacer consultas a métodos propios de bioRuby. Los parámetros son: el archivo de GenBank y el nombre del archivo de salida, donde se guardará el archivo transformado a FASTA.

Luego, se abre el archivo, indicando que el tipo de dato a abrir es GenBank; y luego se itera por cada uno de los valores (a esa variable la seteamos como entry). La variable “definition” es la que le da el nombre de la primer línea (la que comienza con “>”). Luego se consulta por la secuencia (entry.seq) y se pide que haga una traducción (mediante el método translate) a aminoácidos. Esa data la guarda traducida en el formato que corresponde, seteando “definition” al parámetro de entrada de la función to_fasta y se almacena en el archivo seteado en la variable “f”.

- Como correr:

ruby Ejercicio1.rb entrada.gb salida

Ejercicio 2A: BLAST

- Ejercicio realizado en BioRuby, con referencia a la documentación en GitHub de BioRuby (ver link).
- Objetivo: Realizar un script que tenga por entrada el archivo de salida del ejercicio 1, y tenga por salida una lista de las posibles coincidencias en la secuencia de aminoácidos, con otros valores (Accesion, Definition, Número de coincidencias, etc...)
- Código: Nos hemos basado en dos códigos para la realización del mismo: el provisto en GitHub por BioRuby, y de otro usuario (ver link en bibliografía). Hay dos formas de acceder a la Base de Datos de Blast: local o remoto. Para hacerlo local, debe seguir los pasos que se enumeran en el enunciado y luego aplicar el siguiente comando por consola:

makeblastdb -in swissprot -dbtype 'prot' -out swissprot

Por otro lado se puede obviar esta parte y hacer una llamada remota

```
if ARGV[2].eq?('--local')
  blast = Bio::Blast.local('blastp', '/home/kevin/Desktop/ex2/swissprot')
elsif ARGV[2].eq?('--remoto')
  blast = Bio::Blast.remote('blastp', 'swissprot', '-e 0.0001', 'genomenet')
else
  puts 'No es valido'
  exit
end

entries = Bio::FlatFile.open(Bio::FastaFormat, ARGV[0])
```

Recuerde que en caso de querer hacerlo local, deberá cambiar el path que aparece en el script y ejecutar el comando de *makeblastdb*.

Luego se procede a abrir el archivo, indicando por parámetro que es un archivo de formato FASTA, y se itera por cada uno de los hits que se encontraron en la base de datos.

Hemos probado realizar la consulta con distintos outputs (de distintos archivos .gb) obtenidos del Ejercicio 1. Hemos notado un caso en particular, se encuentran 0 hits.

El archivo Ejercicio2_bioruby_docs.rb es un ejemplo base del cual partimos a investigar para hacer las consultas BLAST y para ver que datos podemos visualizar.

Para correr el script debe hacer:

```
ruby Ejercicio2.rb entrada.fas salida <--local | --remoto>
```

Ejercicio 2B: Interpretación del resultado del BLAST

La función BLAST toma la secuencia de aminoácidos original y lo compara con una base de datos, devuelve una lista de todas las secuencias que tienen similitudes con la original.

En el archivo. blast que se obtiene como salida se devuelve cada hit separado con líneas. En donde se puede ver: #numero de hit.

El número de accesoión que el que identifica un registro de secuencia en la base de datos GenBank 5 como también la definición del hit que lo describe.

Luego se observan todos los valores estadísticos asociados con las secuencias encontradas (Serán explicados en la parte 2) como también se puede ver la "Query Sequence" qué es la secuencia original y "Subject Sequence" que es con la que se compara.

Debajo se muestra el código generado en ruby para visualizar el resultado y lo que realmente se obtiene al correr el blast.

```
File.open(ARGV[1], 'w') do |f|
  entries.each_entry do |entry|

    report = blast.query(entry.seq)

    report.hits.each_with_index do |hit, hit_index|
      f.puts '-----'
      f.puts "Hit #{hit_index}"
      f.puts hit.accession
      f.puts hit.definition
      f.puts " - Query length: #{hit.len}"
      f.puts " - Number of identities: #{hit.identity}"
      f.puts " - Length of Overlapping region: #{hit.overlap}"
      f.puts " - % Overlapping: #{hit.percent_identity}"
      f.puts " - Query sequence: #{hit.query_seq}"
      f.puts " - Subject sequence: #{hit.target_seq}"
      hit.hsps.each_with_index do |hsps, hsps_index|
        f.puts " - Bit score: #{hsps.bit_score}"
        f.puts " - Gaps: #{hsps.gaps}"
      end
    end
  end
end
```

```

Hit 0
26435
P11532.3 RecName: Full=Dystrophin
- Query length: 3685
- Number of identities: 29
- Length of Overlapping region: 31
- % Overlapping:
- Query sequence: LQAIEREKA EKFRKLQDASRSAQALVEQMVN
- Subject sequence: VNAIEREKA EKFRKLQDASRSAQALVEQMVN
- Bit score: 59.6918
- Gaps: 0

```

Explicación sobre el significado de los valores estadísticos asociados a las secuencias encontradas

Los valores estadísticos asociados a las secuencias son como se muestra en el resultado del BLAST, el número de identidades, el bit score y el gaps. A continuación se presentan definiciones de cada uno de estos valores, como también la definición de alineamiento para poder entender mejor el proceso.

- **Número de Identidades:** El grado en que dos secuencias (de nucleótidos o aminoácidos) tienen los mismos residuos en las mismas posiciones en una alineación, a menudo expresada con un porcentaje.
- **Alineamiento:** Es el proceso o resultado de emparejar los residuos de nucleótidos o aminoácidos de dos o más secuencias biológicas para lograr niveles máximos de identidad, y en el caso de secuencias de aminoácidos, conservación, con el fin de evaluar el grado de similitud y la posibilidad de homología.
- **Bit score:** El bit score se deriva del puntaje de alineación sin procesar, teniendo en cuenta las propiedades estadísticas del sistema de puntuación. Debido a que los puntajes de bits se normalizan con respecto al sistema de puntuación, se pueden usar para comparar puntaje de alineación de diferentes búsquedas.
- **Gaps:** Es un espacio introducido en una alineación para compensar las inserciones y eliminaciones en una secuencia relativa a otra. Para evitar la acumulación de demasiadas lagunas en una alineación, la introducción de una brecha provoca la deducción de una cantidad fija (la puntuación de la brecha) de la puntuación de alineación. La extensión de la brecha para abarcar nucleótidos o aminoácidos adicionales también se penaliza en la puntuación de una alineación.

Referencia: <https://www.ncbi.nlm.nih.gov/books/NBK62051/>

Ejercicio 3: Multiple Sequence Alignment (MSA)

Para hacer el MSA se utilizó [MUSCLE \(Multiple Sequence Comparison by Log- Expectation\)](#)

Y concluimos que a mayor diferencia entre "Query length", menos alineadas están las secuencias.

Para realizar la experiencia, se creó un archivo BLAST con 500 hits:

1. Se hizo una primera medición MSA donde se compararon las secuencias de los primeros 3 hits
2. Se hizo una segunda medición MSA donde se compararon las secuencias de los hits número 0, 250, 500

Para este caso, hemos utilizado el caso particular de **Ubiquinol-cytochrome-c reductase complex cytochrome b subunit**, (ejemplo de FASTA provisto en Wikipedia, y que hemos utilizado en variadas ocasiones para hacer pruebas). Además, hemos utilizado este ya que al hacer las consultas BLAST, hemos tenido un número de hits de 500 en total (Frente al otro que hemos tomado como referencia -Human muscular dystrophy gene region- que tuvo un hit de 5 en total)

Primera medición

Se usó [la proteína O47885](#).

- Query length: 378
- Number of identities: 251
- Length of Overlapping region: 284
- Bit score: 500.745
- Gaps: 0

```
>sp|O47885.1|CYB_ELEMA RecName: Full=Cytochrome b; AltName: Full=Complex III
subunit 3; AltName: Full=Complex III subunit III; AltName: Full=Cytochrome b-c1
complex subunit 3; AltName: Full=Ubiquinol-cytochrome-c reductase complex
cytochrome b subunit
MTHTRKFHPLFKIINKSFIDLPTPSNISTWWNFGSLLGACLTQILTGLFLAMHYTPDTMTAFSSMSHIC
RDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWGQMSF
WGATVITNLFSAPYIGTNLVEWIIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLG
LTSDSDKIPFHPYYTIKDFLGLLILILLLLLLALLSPDMLGDPDNYPADPLNTPLHIKPEWYFLFAYAI
LRVSPNKLGGVLAFLSILILGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEHPYIII
GQMASILYFSIILAFPLIAGMIENYLIK
```

Se usó [la proteína P24958](#)

- Query length: 378
- Number of identities: 249
- Length of Overlapping region: 284
- Bit score: 477.248
- Gaps: 0

```
>sp|P24958.2|CYB_LOXAF RecName: Full=Cytochrome b; AltName: Full=Complex III
subunit 3; AltName: Full=Complex III subunit III; AltName: Full=Cytochrome b-c1
complex subunit 3; AltName: Full=Ubiquinol-cytochrome-c reductase complex
cytochrome b subunit
MTHIRKSHPLLKIINKSFIDLPTPSNISTWWNFGSLLGACLTQILTGLFLAMHYTPDTMTAFSSMSHICRDVNYGWIIRQLH
SNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWGQMSFWGATVITNLFSAPYIGTNLVEWIIW
GFSVDKATLNRFFALHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLLLLALLSPDM
```


LGDPDNYPADPLNTPHLIKPEWYFLFAYAILRSVPNKLGGVLALLLSILILGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDL
LTLTWIGSQPVEYPYIIIGQMASILYFSIILAFLPIAGVIENYLIK

Se usó [la proteína P92658](#)

- Query length: 378
- Number of identities: 247
- Length of Overlapping region: 284
- Bit score: 475.322
- Gaps: 0

```
>sp|P92658.3|CYB_MAMPR RecName: Full=Cytochrome b; AltName: Full=Complex III
subunit 3; AltName: Full=Complex III subunit III; AltName: Full=Cytochrome b-c1
complex subunit 3; AltName: Full=Ubiquinol-cytochrome-c reductase complex
cytochrome b subunit
MTHIRKSHPLLKILNKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDPTMTAFSSMSHICRDVNYGWIIRQLH
SNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTDLVEWIWG
GFSVDKATLNRFFALHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILFLLLLALLSPDM
LGDPDNYPADPLNTPHLIKPEWYFLFAYAILRSVPNKLGGVLALLLSILILGLMPLLHTSKHRSMMLRPLSQVLFWTLATDL
LMLTWIGSQPVEYPYIIIGQMASILYFSIILAFLPIAGMIENYLIK
```

sp P92658.3 CYB_MAMPR	MTHIRKsHPLLKILNKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDPTM
sp O47885.1 CYB_ELEMA	MTHtRKfHPLfKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDPTM
sp P24958.2 CYB_LOXAF	MTHIRKsHPLLKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDPTM
sp P92658.3 CYB_MAMPR	TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL
sp O47885.1 CYB_ELEMA	TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL
sp P24958.2 CYB_LOXAF	TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL
sp P92658.3 CYB_MAMPR	LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTdLVEWIWGGFSVDKATLNRFFA
sp O47885.1 CYB_ELEMA	LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFA
sp P24958.2 CYB_LOXAF	LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFA
sp P92658.3 CYB_MAMPR	LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILfLL
sp O47885.1 CYB_ELEMA	fHFILPFTMvALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL
sp P24958.2 CYB_LOXAF	LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL
sp P92658.3 CYB_MAMPR	LLALLSPDMLGDPDNYPADPLNTPHLIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI
sp O47885.1 CYB_ELEMA	LLALLSPDMLGDPDNYPADPLNTPHLIKPEWYFLFAYAILRSVPNKLGGVLALfLSILI
sp P24958.2 CYB_LOXAF	LLALLSPDMLGDPDNYPADPLNTPHLIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI
sp P92658.3 CYB_MAMPR	LGiMPLLHTSKHRSMMLRPLSQVLFWTLatDLLmLTWIGSQPVEYPYIIIGQMASILYFS
sp O47885.1 CYB_ELEMA	LGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEhPYIIIGQMASILYFS
sp P24958.2 CYB_LOXAF	LGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEYPYIIIGQMASILYFS
sp P92658.3 CYB_MAMPR	IILAFLPIAGMIENYLIK
sp O47885.1 CYB_ELEMA	IILAFLPIAGMIENYLIK
sp P24958.2 CYB_LOXAF	IILAFLPIAGvIENYLIK

Segunda medición

Se usó [la proteína O47885](#).

- Query length: 378
- Number of identities: 251
- Length of Overlapping region: 284
- Bit score: 500.745

- Gaps: 0

```
>sp|O47885.1|CYB_ELEMA RecName: Full=Cytochrome b; AltName: Full=Complex III
subunit 3; AltName: Full=Complex III subunit III; AltName: Full=Cytochrome b-c1
complex subunit 3; AltName: Full=Ubiquinol-cytochrome-c reductase complex
cytochrome b subunit
MTHTRKFHPLFKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTMTAFSSMSHIC
RDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWQMSF
WGATVITNLFSAPYIGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLG
LTSDSDKIPFHPYYTIKDFLGLLILILLLLLLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAI
LRSVPNKLGGVLAFLSILILGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEHPYIII
GQMASILYFSIILAFLPIAGMIENYLIK
```

Se usó [la proteína Q9T4R0](#)

- Query length: 380
- Number of identities: 201
- Length of Overlapping region: 283
- Bit score: 416.387
- Gaps: 0

```
>sp|Q9T4R0.1|CYB_ELIMA RecName: Full=Cytochrome b; AltName: Full=Complex III
subunit 3; AltName: Full=Complex III subunit III; AltName: Full=Cytochrome b-c1
complex subunit 3; AltName: Full=Ubiquinol-cytochrome-c reductase complex
cytochrome b subunit
MTNIRKSHPLLKIINHNSFIDLPTPSNISSWWNFGSLLGICLILQIATGLFLAMHYTSDTTTAFSSVTHIC
RDVNYGWLIRYLHANGASMFICLFHVGGRGMYYSIETWNMGIIILFAVMATAFMGYVLPWQMSF
WGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLTRFFAFHFILPFIIVALVMVHLLFLHETGSNNPSG
LNSDADKIPFHPYYTIKDILGVLLFLFLISLVLFAPDLLGDPDNYTPANPLNTPPHIKPEWYFLFAYAI
LRSIPNKLGGVLAFLSILVLALIPHLHTSKLQSLMFRPLTQALYWILVADLLILTWIGGQPVEYFPIII
GQLASVLYFAIILIFMPMAGMIEDSILKMD
```

Se usó [la proteína Q7JE02](#):

- Query length: 379
- Number of identities: 203
- Length of Overlapping region: 283
- Bit score: 407.142
- Gaps: 0

```
>sp|Q7JE02.1|CYB_MUSEV RecName: Full=Cytochrome b; AltName: Full=Complex III
subunit 3; AltName: Full=Complex III subunit III; AltName: Full=Cytochrome b-c1
complex subunit 3; AltName: Full=Ubiquinol-cytochrome-c reductase complex
cytochrome b subunit
MTNIRKTHPLTKIINNSFIDLPAAPSNIASWWNFGSLLGICLIQILTGLFLAMHYTSDTATAFSSVTHIC
RDVNYGWIIRYMHANGASMFICLFHVGRLYYGSYMTETWNIGIILLFAVMATAFMGYVLPWQMSF
WGATVITNLLSAIPYIGTNLVEWIWGGFSVDKATLTRFFAFHFILPFIISALAAVHLLFLHETGSNNPSG
IPSDSDKIPFHPYYTIKDILGALLLILMLTLLVLFSPDLLGDPDNYIPANPLNTPPHIKPEWYFLFAYAI
LRSIPNKLGGVLAFLSILILAIIPLLHTSKQRSMMFRPLSQCLFWLLVADLLTLTWIGGQPVEHPFIII
GQLASILYFMILLVLMPIISIIENMLKW
```

sp	Q047885.1	CYB_ELEMA	MThtrRKfHPLfKIInkSFIDLPTSPNISWNNFSGLLGaCLItQILtGLFLAMHYTpDtm
sp	Q9T4R0.1	CYB_ELIMA	MTNIRKsHPLlKIInhSFIDLPTSPNISWNNFSGLLGIClIiQIaTGLFLAMHYTSdt
sp	Q7JE02.1	CYB_MUSEV	MTNIRKtHPLtKIInnSFIDLPaPSNISAWNNFSGLLGIClIiQILtGLFLAMHYTSdTa
sp	Q047885.1	CYB_ELEMA	TAFSSmsHICRDVNYGWIIRqLHsNGASiFFlCLytHiGrnIYYGSYlysETWNtGImLI
sp	Q9T4R0.1	CYB_ELIMA	TAFSSVTHICRDVNYGWLIRYLHANGASMFFIClFihVGRGmYYGSYMsietWNmGIILL
sp	Q7JE02.1	CYB_MUSEV	TAFSSVTHICRDVNYGWIIRYmHANGASMFFIClFhVGRGLYYGSYMfetETWniGIILL
sp	Q047885.1	CYB_ELEMA	litMATAFMGYVLPWGQMSFWGATVITNlfSAIPYIGTNLVEWIWGGFSVDKATlnRFFA
sp	Q9T4R0.1	CYB_ELIMA	FAVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTtLVEWIWGGFSVDKATlTRFFA
sp	Q7JE02.1	CYB_MUSEV	FAVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTNLVEWIWGGFSVDKATlTRFFA
sp	Q047885.1	CYB_ELEMA	FHFILPFtmvALAgVHLtFLHETGSNNPlGLtSDSDKIPFHPYYTIKIdfLGlLiLiLLlLl
sp	Q9T4R0.1	CYB_ELIMA	FHFILPFIiivALvmVHLLFLHETGSNNPsGLnSDaDKIPFHPYYTIKIDlGvLLLfLFLi
sp	Q7JE02.1	CYB_MUSEV	FHFILPFIiisALaaVHLLFLHETGSNNPsGipSDSDKIPFHPYYTIKIDlGaLLLlILmLt
sp	Q047885.1	CYB_ELEMA	lLaIlSPDmLGDPDNYmPadPLNTPlHIKPEWYFLFAYAILRSvPNKLGGVLALflSILI
sp	Q9T4R0.1	CYB_ELIMA	sLVLFaPDLLGDPDNYtPANPLNTPPHIKPEWYFLFAYAILRSIPNKLGGVLALILSILv
sp	Q7JE02.1	CYB_MUSEV	lLVLFSPDLLGDPDNYiPANPLNTPPHIKPEWYFLFAYAILRSIPNKLGGVLALIfSILI
sp	Q047885.1	CYB_ELEMA	LgLmPlLHTSKhRSMMlRPLSQvLFWtLtmDLLTLTWIGsQPVEHPyIIIGQmASILYfs
sp	Q9T4R0.1	CYB_ELIMA	LALIPhLHTSKlqSLmFRPltQaLyWiLVADLLiLTWIGGQPVEYpFIIIGQLASvLYFa
sp	Q7JE02.1	CYB_MUSEV	LaiiPlLHTSKqRSMMFRPLSQcLFWlLVADLLTLTWIGGQPVEHPFIIIGQLASILYfm
sp	Q047885.1	CYB_ELEMA	IILAFIPIAGMIENyLiK--
sp	Q9T4R0.1	CYB_ELIMA	IILiFMPmAGMIEdsILKmd
sp	Q7JE02.1	CYB_MUSEV	IlLvIMPIisiIEInnmLKw-

Referencias bibliográficas

https://www.ncbi.nlm.nih.gov/nuccore/NM_000531.5

<http://bioruby.org/rdoc/Bio/Sequence.html>

<https://github.com/bioruby/bioruby/blob/master/doc/Tutorial.rd>

<https://gist.github.com/jimhester/3169859>

http://www.cbcb.umd.edu/confcour/Fall2012/Bio_UtilsDBProject_2012.pdf

<https://github.com/bioruby/documents/wiki/SampleCodes>

<http://bioruby.org/rdoc/Bio/GenBank.html>

https://es.wikipedia.org/wiki/Marco_abierto_de_lectura