



# TP1: Introducción a la Bioinformática

Alumnos:

- Kevin Cortes
- Daniel Lobo
- Magdalena Vega



## 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.



# Ejercicio 1: Procesamiento de secuencias

## Genbank:

- Encabezamiento
- Referencias bibliográficas
- Tabla de características
- Secuencias

## FASTA:

- Formato basado en texto
- Representa secuencias de ácidos nucleicos, péptido, o aminoácidos

# Ejercicio 1: Procesamiento de secuencias

## Secuencia de aminoácidos:

- Recibe como input el archivo Genbank.

```
#!/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
```




## Ejercicio 2A: BLAST

- 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...)
- Uso de BioRuby

## Ejercicio 2A: BLAST

```
if ARGV[2].eq?('--local')
  blast = Bio::Blast.local('blastp', '/home/kevin/Desktop/ex2/swissprot')
elsif ARGV[2].eq?('--remote')
  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])
```



## Ejercicio 2B: Interpretación del resultado del BLAST

- 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.

## Ejercicio 2B: Interpretación del resultado del 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: LQAIEREKAEKFRKLQDASRSAQALVEQMVN
- Subject sequence: VNAIEREKAEEKFRKLQDASRSAQALVEQMVN
- Bit score: 59.6918
- Gaps: 0
```





## Ejercicio 3: Multiple Sequence Alignment (MSA)

- <https://www.ebi.ac.uk/Tools/msa/muscle/>
  - MUSCLE stands for MULTiple Sequence Comparison by Log-Expectation.
- A mayor diferencia entre “Query length”, menos alineadas están las secuencias.
- Se creó un archivo BLAST con 500 hits:
  - 1) Se compararon las secuencias de los primeros 3 hits
  - 2) Se compararon las secuencias de los hits número 0, 250, 500



## Ejercicio 3: Multiple Sequence Alignment (MSA)

- <https://www.ncbi.nlm.nih.gov/protein/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

MTHTRKHFHPLFKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTMTAFSSMSHIC  
RDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWQGQMSF  
WGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLG  
LTSDSDKIPFHPYYTIKDFLGLLILILLLLLLALLSPDMLGDPDNYMPADPLNTPHLIKPEWYFLFAYAI  
LRSVPNKLGGVLALFLSILILGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEHPYIII  
GQMASILYFSIILAFPLIAGMIENYLIK



## Ejercicio 3: Multiple Sequence Alignment (MSA)

- <https://www.ncbi.nlm.nih.gov/protein/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

MTHIRKSHPLLKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTMTAFSSMSHICRDVNYGWIIRQLHSNGA  
SIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKA  
TLNRFFALHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILLLLLLLALLSPDMLGDPDNYPADPL  
NTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLSILILGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEYPYIIIG  
QMASILYFSIILAFLPIAGVIENYLIK



## Ejercicio 3: Multiple Sequence Alignment (MSA)

- <https://www.ncbi.nlm.nih.gov/protein/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

MTHIRKSHPLLKILNKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTMTAFSSMSHICRDVNYGWIIRQLHSNGA  
SIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWGQMSFWGATVITNLFSaipYIGTDLVEWiwGGFSVDKA  
TLNRFFALHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILFLLLLALLSPDMLGDPDNYMPADPL  
NTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSILILGIMPLLHTSKHRSMMMLRPLSQVLFWTLATDLLMLTWIGSQPVEYPYIIIG  
QMASILYFSIILAFLPIAGMIENYLIK

sp | P92658.3 | CYB\_MAMPR  
sp | O47885.1 | CYB\_ELEMA  
sp | P24958.2 | CYB\_LOXAF

MTHIRKsHPLLKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM  
MTHtRKfHPLfKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM  
MTHIRKsHPLLKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM

sp | P92658.3 | CYB\_MAMPR  
sp | O47885.1 | CYB\_ELEMA  
sp | P24958.2 | CYB\_LOXAF

TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL  
TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL  
TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL

sp | P92658.3 | CYB\_MAMPR  
sp | O47885.1 | CYB\_ELEMA  
sp | P24958.2 | CYB\_LOXAF

LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTdLVEWIWGGFSVDKATLNRFFA  
LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFA  
LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFA

sp | P92658.3 | CYB\_MAMPR  
sp | O47885.1 | CYB\_ELEMA  
sp | P24958.2 | CYB\_LOXAF

LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILfLL  
fHFILPFTMvALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL  
LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL

sp | P92658.3 | CYB\_MAMPR  
sp | O47885.1 | CYB\_ELEMA  
sp | P24958.2 | CYB\_LOXAF

LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI  
LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALfLSILI  
LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI

sp | P92658.3 | CYB\_MAMPR  
sp | O47885.1 | CYB\_ELEMA  
sp | P24958.2 | CYB\_LOXAF

LGiMPLLHTSKHRSMMLRPLSQVLFWTlatDLLmLTWIGSQPVEYPYIIIGQMASILYFS  
LGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEhPYIIIGQMASILYFS  
LGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEYPYIIIGQMASILYFS

sp | P92658.3 | CYB\_MAMPR  
sp | O47885.1 | CYB\_ELEMA  
sp | P24958.2 | CYB\_LOXAF

IILAFLPIAGMIENYLIK  
IILAFLPIAGMIENYLIK  
IILAFLPIAGvIENYLIK




## Ejercicio 3: Multiple Sequence Alignment (MSA)

- <https://www.ncbi.nlm.nih.gov/protein/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  
RDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWQGMSF  
WGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLG  
LTSDSDKIPFHPYYTIKDFLGLLILLLLLLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAI  
LRSVPNKLGGVLALFLSILILGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEHPYIII  
GQMASILYFSIILAFPLIAGMIENYLIK



## Ejercicio 3: Multiple Sequence Alignment (MSA)

- <https://www.ncbi.nlm.nih.gov/protein/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

MTNIRKSHPLLKIINHSFIDLPTPSNISSWWNFGSLLGICLILQIATGLFLAMHYTSDTTTAFSSVTHIC  
RDVNYGWLIRYLHANGASMFFICLFHVGRGMYGYSMSIETWNMGIILLFAVMATAFMGYVLPWGQMSF  
WGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLTRFFAFHFILPFIIVALVMVHLLFLHETGSNNPSG  
LNSDADKIPFHPYYTIKDILGVLLLFLFLISLVLFAPDLLGDPDNYTPANPLNTPPHIKPEWYFLFAYAI  
LRSIPNKLGGVLALILSILVLALIPHLHTSKLQSLMFRPLTQALYWILVADLLILTWIGGQPVEYPFIII  
GQLASVLYFAIILIFMPMAGMIEDSILKMD



## Ejercicio 3: Multiple Sequence Alignment (MSA)

- <https://www.ncbi.nlm.nih.gov/protein/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

```
MTNIRKTHPLTKIINNSFIDLPAASNISAWWNFGSLLGICLIQILTGLFLAMHYTSDTATAFSSVTHIC
RDVNYGWIIRYMHANGASMFFICLFLHVGRGLYYGSYMTETWNIGIILLFAVMATAFMGYVLPWGQMSF
WGATVITNLLSAIPYIGTNLVEWIWGGFSVDKATLTRFFAFHFILPFIISALAAVHLLFLHETGSNNPSG
IPSDSDKIPFHPYYTIKDILGALLLILMLTLLVLFSPDLLGDPDNYIPANPLNTPPHIKPEWYFLFAYAI
LRSIPNKLGGVLALIFSILIAIPLLHTSKQRSMMFRPLSQCLFWLLVADLLTLTWIGGQPVEHPFIII
GQLASILYFMILLVLMPIISIIENNMLKW
```



sp		O47885.1		CYB_ELEMA		MTh	tRKf	HPLf	KIINK	SFIDL	TPSNIS	tWW	NFGS	LLGa	CLIt	QILT	TGLF	LAMHY	Tp	DTm						
sp		Q9T4R0.1		CYB_ELIMA		MTN	IRKs	HPLl	KIINH	SFIDL	TPSNIS	SSWW	NFGS	LLGIC	LIl	QIa	TGLF	LAMHY	TSDT	t						
sp		Q7JE02.1		CYB_MUSEV		MTN	IRKt	HPLt	KIINN	SFIDL	PaPS	NISAW	WNFG	SLLG	ICLI	liQ	ILT	TGLF	LAMHY	TSDTa						
sp		O47885.1		CYB_ELEMA		TAF	SSms	HICR	DVNY	GWII	RqLH	sNGA	SiFF	lCL	ytHi	GRnI	YYGS	Ylys	ETWN	tGIm	LL					
sp		Q9T4R0.1		CYB_ELIMA		TAF	SSVTH	ICR	DVNY	GWl	IRYL	HANG	ASMFF	ICLF	iHV	GRGm	YYGS	SYMsi	ETWN	mGI	ILL					
sp		Q7JE02.1		CYB_MUSEV		TAF	SSVTH	ICR	DVNY	GWII	RYmH	ANGA	SMFF	ICLF	lHV	GRGL	YYGS	SYMft	ETWN	iGI	ILL					
sp		O47885.1		CYB_ELEMA		lit	MATA	FMG	YVLP	WQG	MSFW	GATV	ITNL	fSA	IPYI	GTNL	VEWI	WGGS	VDKAT	Lnr	FFA					
sp		Q9T4R0.1		CYB_ELIMA		FAV	MATA	FMG	YVLP	WQG	MSFW	GATV	ITNL	LSA	IPYI	GTt	LVEW	IWGG	FSDKAT	LTR	FFA					
sp		Q7JE02.1		CYB_MUSEV		FAV	MATA	FMG	YVLP	WQG	MSFW	GATV	ITNL	LSA	IPYI	GTNL	VEWI	WGGS	VDKAT	LTR	FFA					
sp		O47885.1		CYB_ELEMA		FHF	ILPF	tmv	ALAg	VHLt	FLH	ETGS	NNPl	GLt	SDS	DKIP	FHPY	YTIK	DfL	Gll	LiLi	LlLl	Ll			
sp		Q9T4R0.1		CYB_ELIMA		FHF	ILPF	IIv	ALvm	VHLL	FLH	ETGS	NNPs	GLn	SDa	DKIP	FHPY	YTIK	DIL	Gv	LLL	fLFL	Li			
sp		Q7JE02.1		CYB_MUSEV		FHF	ILPF	IIIs	ALAa	VHLL	FLH	ETGS	NNPs	Gip	SDS	DKIP	FHPY	YTIK	DIL	Ga	LLL	LiLm	Lt			
sp		O47885.1		CYB_ELEMA		lLa	Ll	SPD	mLGD	PDNY	mPa	dPL	NTPl	hIK	PEWY	FLF	AYAIL	RSv	PNKL	GGV	LAL	fLS	SILI			
sp		Q9T4R0.1		CYB_ELIMA		sLV	LFa	PDLL	GDP	NYt	PAN	PLNT	PPH	IK	PEWY	FLF	AYAIL	RSIP	PNKL	GGV	LAL	ILS	SILv			
sp		Q7JE02.1		CYB_MUSEV		lLV	LFSP	DLL	GDP	NYi	PAN	PLNT	PPH	IK	PEWY	FLF	AYAIL	RSIP	PNKL	GGV	LAL	ifS	SILI			
sp		O47885.1		CYB_ELEMA		Lg	Lm	Pll	LHT	SKh	RSM	Mr	PLS	Qv	LFw	tLtm	DLL	TLT	WIG	sQP	VEHP	YII	IGQ	mAS	ILY	Fs
sp		Q9T4R0.1		CYB_ELIMA		LAL	IPh	LHT	SKl	qSl	MFR	PLt	Qa	Ly	Wi	LVAD	LLi	LTW	IGG	QP	VEy	PFII	IGQ	LAS	vLY	Fa
sp		Q7JE02.1		CYB_MUSEV		LAi	iPl	LHT	SKq	RSM	MFR	PLS	Qc	LFw	lLVAD	LLTLT	WIGG	QP	VEHP	FI	II	IGQ	LAS	ILY	Fm	
sp		O47885.1		CYB_ELEMA		II	LA	Fl	PIA	GMI	ENy	Li	K--													
sp		Q9T4R0.1		CYB_ELIMA		II	Li	FMP	mAG	MI	Eds	IL	Kmd													
sp		Q7JE02.1		CYB_MUSEV		Il	Lv	lMP	Iisi	IE	Nnm	LKw-														