UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo
Unidad Iztapalapa

Bioinformática aplicada a recursos genéticos

Dr. Miguel Ángel del Río Portilla
Profesor Investigador, Cátedra "Alfonso Villalobos"

Dra. Irene de los Ángeles Barriga Sosa
Profesor Investigador tiempo completo

Dra. Dra. Erika Magallón Gayón
Posdoctorado

CICESE

1

Organizadores

- Dra. Irene de los Ángeles Barriga Sosa
- Dr. Miguel Ángel del Río Portilla
- Departamento de Hidrobiología
- UAM- Iztapalapa

2

Bioinformática

* Colección, clasificación, almacenamiento y análisis de información bioquímica y biológica utilizando computadoras, especialmente con aplicación a genética molecular y genómica. (https://www.merriam-webster.com/dictionary/bioinformatics)
* La suma de aproximaciones computacionales para analizar, manejar y almacenar datos biológicos

3

Código Genético

* El conjunto de correspondencia entre tripletes (tres nucleótidos) de DNA y los aminoácidos en proteínas(1)
* Anteriormente se pensaba que era estático (2)

Códigos genéticos considerados en el NCBi (3)

1. The Standard Code
2. The Vertebrate Mitochondrial Code
3. The Yeast Mitochondrial Code
4. The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code
5. The Invertebrate Mitochondrial Code
6. The Ciliate, Dasycladacean and Hexamita Nuclear Code

9. The Echinoderm and Flatworm Mitochondrial Code
10. The Euplotid Nuclear Code
11. The Bacterial, Archaeal and Plant Plastid Code
12. The Alternative Yeast Nuclear Code
13. The Ascidian Mitochondrial Code
14. The Alternative Flatworm Mitochondrial Code
16. Chlorophycean Mitochondrial Code
21. Trematode Mitochondrial Code
22. Scenedesmus obliquus Mitochondrial Code
23. Thraustochytrium Mitochondrial Code
24. Pterobranchia Mitochondrial Code
25. Candidate Division SR1 and Gracilibacteria Code
26. Pachysolen tannophilus Nuclear Code
27. Karyorelict Nuclear
28. Condylostoma Nuclear
29. Mesodinium Nuclear
30. Peritrich Nuclear
31. Blastocrithidia Nuclear

1. https://www.ncbi.nlm.nih.gov/books/NBK21950/
2. https://www.ncbi.nlm.nih.gov/pubmed/1579111?dopt=Abstract
3. https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

4

Código estándar (1)

TTT	F	Phe	TCT	S	Ser	TAT	Y	Tyr	TGT	C	Cys
TTC	F	Phe	TCC	S	Ser	TAC	Y	Tyr	TGC	C	Cys
TTA	L	Leu	TCA	S	Ser	TAA	*	Ter	TGA	*	Ter
TTG	L	Leu i	TCG	S	Ser	TAG	*	Ter	TGG	W	Trp
CTT	L	Leu	CCT	P	Pro	CAT	H	His	CGT	R	Arg
CTC	L	Leu	CCC	P	Pro	CAC	H	His	CGC	R	Arg
CTA	L	Leu	CCA	P	Pro	CAA	Q	Gln	CGA	R	Arg
CTG	L	Leu i	CCG	P	Pro	CAG	Q	Gln	CGG	R	Arg
ATT	I	Ile	ACT	T	Thr	AAT	N	Asn	AGT	S	Ser
ATC	I	Ile	ACC	T	Thr	AAC	N	Asn	AGC	S	Ser
ATA	I	Ile	ACA	T	Thr	AAA	K	Lys	AGA	R	Arg
ATG	M	Met i	ACG	T	Thr	AAG	K	Lys	AGG	R	Arg
GTT	V	Val	GCT	A	Ala	GAT	D	Asp	GGT	G	Gly
GTC	V	Val	GCC	A	Ala	GAC	D	Asp	GGC	G	Gly
GTA	V	Val	GCA	A	Ala	GAA	E	Glu	GGA	G	Gly
GTG	V	Val	GCG	A	Ala	GAG	E	Glu	GGG	G	Gly

5

Secuenciación (Sanger)

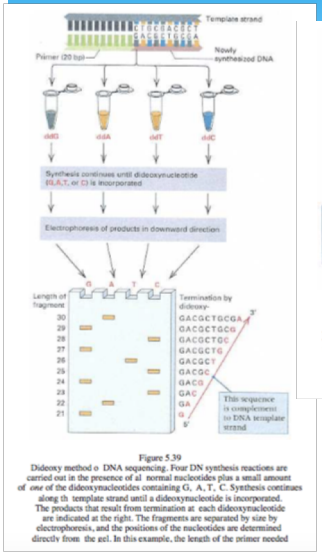


Figure 5.39
Dideoxy method o DNA sequencing. Four DN synthesis reactions are carried out in the presence of al normal nucleotides plus a small amount of one of the dideoxynucleotides containing G, A, T, C. Synthesis continues along th template strand until a dideoxynucleotide is incorporated. The products that result from termination at each dideoxynucleotide are indicated at the right. The fragments are separated by size by electrophoresis, and the positions of the nucleotides are determined directly from the gel. In this example, the length of the primer needed

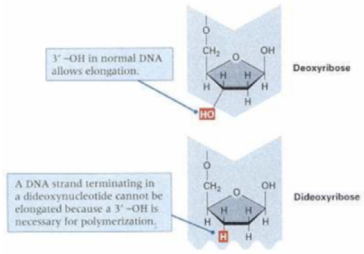


Figure 5.38
Structures of normal deoxyribose and the dideoxyribose sugar used in DNA sequencing. The dideoxyribose has a hydrogen atom (red) attached to the 3′ carbon, in contrast with the hydroxyl group (red) at this position in deoxyribose. Because the 3′ hydroxyl group is essential for the attachment of the next nucleotide in line in a growing DNA strand, the incorporation of a dideoxynucleotide immediately terminates synthesis.

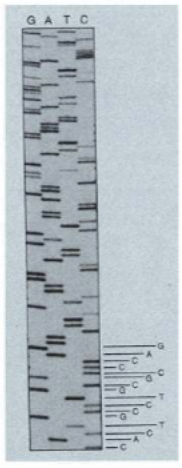


Figure 5.40
A section of a dideoxy sequencing gel. The sequence is read from the bottom to the top. Each horizontal row represents a single nucleotide position in the DNA strand synthesized from the template. The vertical columns result from termination by the dideoxy forms of G, A, T, or C. The sequence from the lower part of the gel is indicated.

6

Hartl, D. L., & Jones, E. W. (1998). *Genetics: Principles and analysis*. Jones and Barlett Publishers.

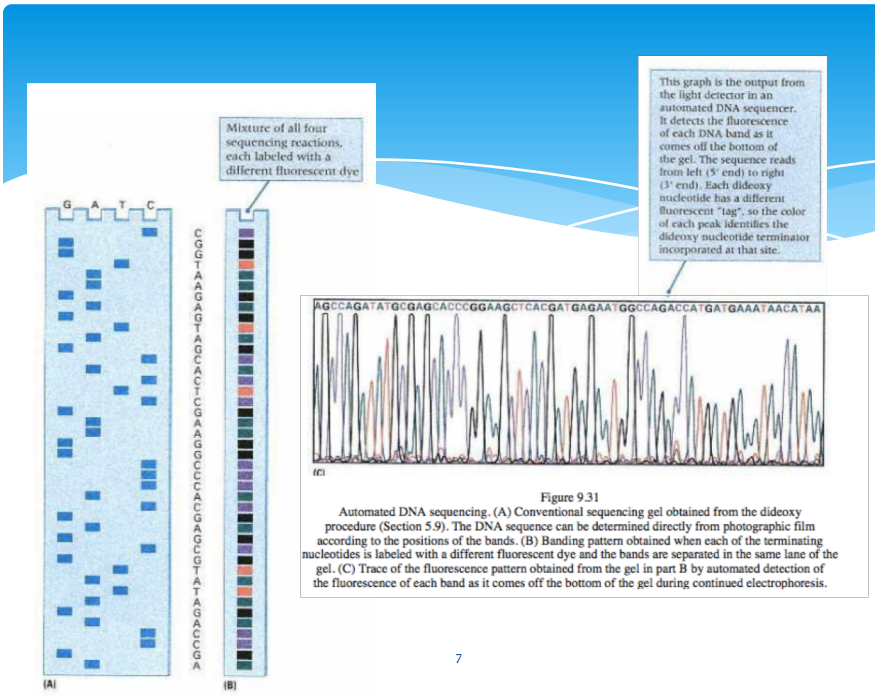


Figure 9.31
Automated DNA sequencing. (A) Conventional sequencing gel obtained from the dideoxy procedure (Section 5.9). The DNA sequence can be determined directly from photographic film according to the positions of the bands. (B) Banding pattern obtained when each of the terminating nucleotides is labeled with a different fluorescent dye and the bands are separated in the same lane of the gel. (C) Trace of the fluorescence pattern obtained from the gel in part B by automated detection of the fluorescence of each band as it comes off the bottom of the gel during continued electrophoresis.

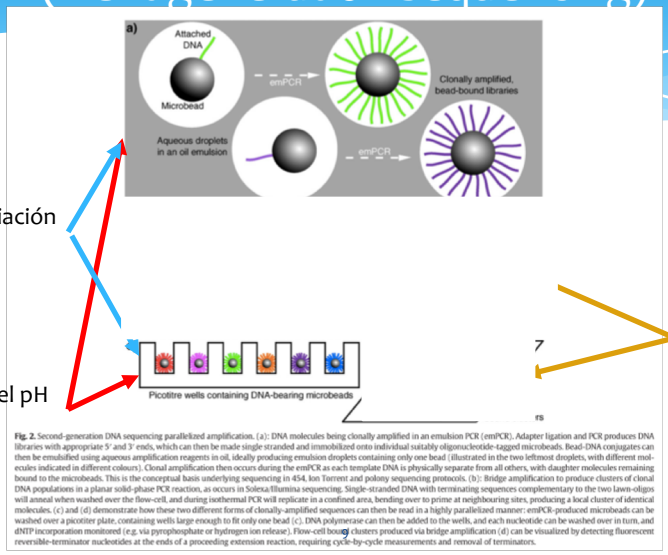
7

Siguiente generación Herramientas Genómicas

- Proyecto del genoma Humano (1990-2003) (~3 Gb)
- Secuenciación Sanger (600-800 pb)
- Desarrollo de nuevas tecnologias
- Next generation Sequencing
- Secuencies de varios millones de fragmentos (lecturas) en una corrida
  - Microsatélites
  - SNP
  - Genomas Mitocondriales

Indispensable para microsatélites y SNPs y análisis de RNA (RNAseq, miRNA…)

Secuenciación de siguiente generación (Next generation sequencing)



Pirosecuenciación 454

Cambios en el pH IonTorrent

Illumina

Fig. 2. Second-generation DNA sequencing parallelized amplification. (a): DNA molecules being clonally amplified in an emulsion PCR (emPCR). Adapter ligation and PCR produces DNA libraries with appropriate 5′ and 3′ ends, which can then be made single stranded and immobilized onto individual suitably oligonucleotide-tagged microbeads. Bead–DNA conjugates can then be emulsified using aqueous amplification reagents in oil, ideally producing emulsion droplets containing only one bead (illustrated in the two leftmost droplets, with different molecules indicated in different colours). Clonal amplification then occurs during the emPCR as each template DNA is physically separate from all others, with daughter molecules remaining bound to the microbeads. This is the conceptual basis underlying sequencing in 454, Ion Torrent and polony sequencing protocols. (b): Bridge amplification to produce clusters of clonal DNA populations in a planar solid-phase PCR reaction, as occurs in Solexa/Illumina sequencing. Single-stranded DNA with terminating sequences complementary to the two lawn-oligos will anneal when washed over the flow-cell, and during isothermal PCR will replicate in a confined area, bending over to prime at neighbouring sites, producing a local cluster of identical molecules. (c) and (d) demonstrate how these two different forms of clonally-amplified sequences can then be read in a highly parallelized manner: emPCR-produced microbeads can be washed over a picotiter plate, containing wells large enough to fit only one bead (c). DNA polymerase can then be added to the wells, and each nucleotide can be washed over in turn, and dNTP incorporation monitored (e.g. via pyrophosphate or hydrogen ion release). Flow-cell bound clusters produced via bridge amplification (d) can be visualized by detecting fluorescent reversible-terminator nucleotides at the ends of a proceeding extension reaction, requiring cycle-by-cycle measurements and removal of terminators.

Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. doi:10.1016/j.ygeno.2015.11.003

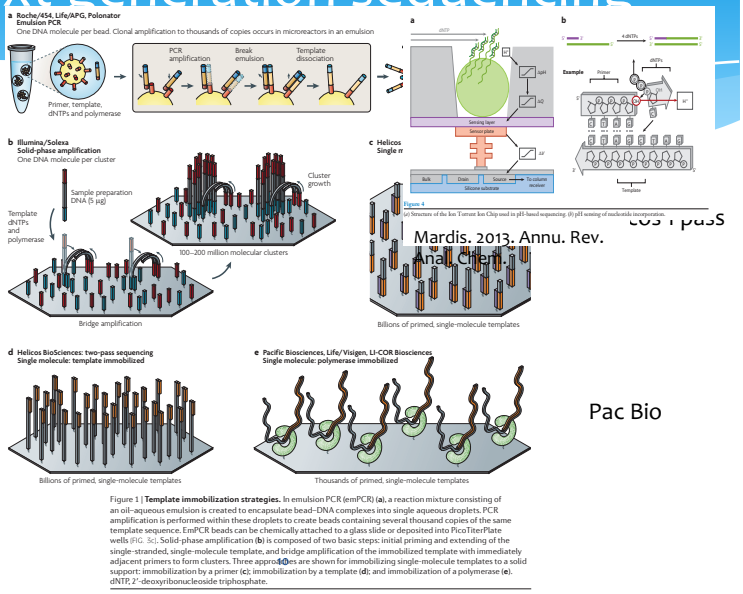
Next generation sequencing

Roche 454

illumina

Helicos 2 pass

Pac Bio



Mardis. 2013. Annu. Rev. Anal. Chem.

Figure 1 | **Template immobilization strategies.** In emulsion PCR (emPCR) (**a**), a reaction mixture consisting of an oil–aqueous emulsion is created to encapsulate bead–DNA complexes into single aqueous droplets. PCR amplification is performed within these droplets to create beads containing several thousand copies of the same template sequence. EmPCR beads can be chemically attached to a glass slide or deposited into PicoTiterPlate wells (FIG. 3c). Solid-phase amplification (**b**) is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. Three approaches are shown for immobilizing single-molecule templates to a solid support: immobilization by a primer (**c**); immobilization by a template (**d**); and immobilization of a polymerase (**e**). dNTP, 2′-deoxyribonucleoside triphosphate.

Metzker 2010. Nature Reviews.

Tercer generación de secuenciación

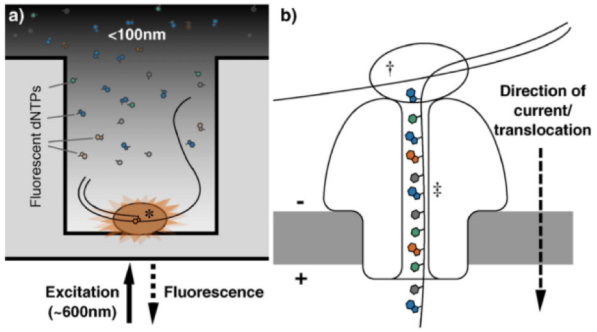


Fig. 3. Third-generation DNA sequencing nucleotide detection. (a): Nucleotide detection in a zero-mode waveguide (ZMW), as featured in PacBio sequencers. DNA polymerase molecules are attached to the bottom of each ZMW (*), and target DNA and fluorescent nucleotides are added. As the diameter is narrower than the excitation light's wavelength, illumination rapidly decays travelling up the ZMW: nucleotides being incorporated during polymerisation at the base of the ZMW provide real-time bursts of fluorescent signal, without undue interference from other labelled dNTPs in solution. (b): Nanopore DNA sequencing as employed in ONT's MinION sequencer. Double stranded DNA gets denatured by a processive enzyme (†) which ratchets one of the strands through a biological nanopore (‡) embedded in a synthetic membrane, across which a voltage is applied. As the ssDNA passes through the nanopore the different bases prevent ionic flow in a distinctive manner, allowing the sequence of the molecule to be inferred by monitoring the current at each channel.

11

Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. doi:10.1016/j.ygeno.2015.11.003

Comparación

12

Análisis de secuencias (Sanger)

Procedimiento

- Toma de muestra
  - Tejido: puede variar de una especie a otra (aleta, branquia, músculo, etc.)
  - Conservación (etanol 95%, congelación)
- Extracción
  - Kit (Qiagen, Promega, etc.)
  - Precipitación con sales (LiCl)
- Amplificación
  - Dirigida al fragmento de interés Iniciadores especificos
- Corroboración de la amplificación
- Limpieza del fragmento
  - Kits
- Envío de la muestra
  - Langebio
  - Macrogen
  - Secuenciador propio

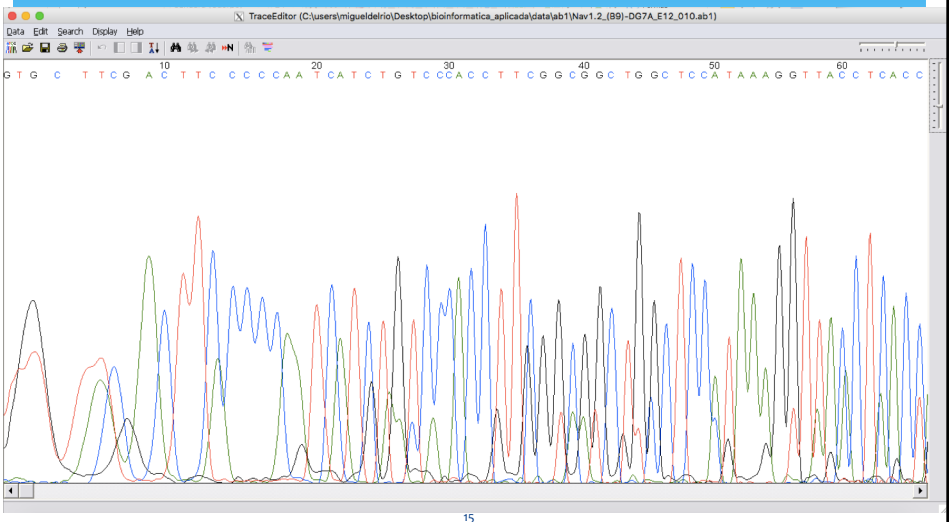
13

Archivo

- AB1 formato AB
- phd.1 formato Phred, (no todos los proveedores lo proporcionan)
- pdf Electroferograma (cromatograma)
- Seq, txt texto

16S_RO_1F-16AR_A07_001.ab1	26/05/2017 11:20	236 KB
16S_RO_1F-16AR_A07_001.seq	26/05/2017 11:20	611 bytes
16S_RO_1R-16BR_A08_002.ab1	26/05/2017 11:20	236 KB
16S_RO_1R-16BR_A08_002.seq	26/05/2017 11:20	613 bytes
16S_RO_2F-16AR_B07_003.ab1	26/05/2017 11:20	236 KB
16S_RO_2F-16AR_B07_003.seq	26/05/2017 11:20	615 bytes
P1AA_16s_16sF.ab1	21/12/2017 22:48	337 KB
P1AA_16s_16sF.pdf	21/12/2017 22:49	43 KB
P1AA_16s_16sF.phd.1	21/12/2017 22:48	6 KB
P1AA_16s_16sF.txt	21/12/2017 22:49	602 bytes
P1AA_16s_16sR.ab1	21/12/2017 22:48	355 KB
P1AA_16s_16sR.pdf	21/12/2017 22:48	41 KB

ab1



15

phd

P1AA_16s_16sF.phd.1

```
BEGIN_SEQUENCE 171221-009_014_P1AA_16s_16sF

BEGIN_COMMENT

CHROMAT_FILE: 171221-009_014_P1AA_16s_16sF
BASECALLER_VERSION: KB 1.4.0
TRACE_PROCESSOR_VERSION: KB 1.4.0
QUALITY_LEVELS: 99
TIME: Thu Dec 21 22:47:32 2017
TRACE_ARRAY_MIN_INDEX: 0
TRACE_ARRAY_MAX_INDEX: 6549
TRIM: -1 -1 -1.000000e+000
TRACE_PEAK_AREA_RATIO: -1.000000e+000
CHEM: term
DYE: big

END_COMMENT

BEGIN_DNA
G 3 3
C 2 18
C 2 25
C 5 34
T 3 44
C 5 57
T 4 73
G 9 82
C 5 91
A 5 99
T 6 115
C 5 127
C 5 136
C 5 148
C 5 158
C 4 170
```

- BEGIN_SEQUENCE 171221-009_014_P1AA_16s_16sF
- BEGIN_COMMENT
- CHROMAT_FILE: 171221-009_014_P1AA_16s_16sF
- BASECALLER_VERSION: KB 1.4.0
- TRACE_PROCESSOR_VERSION: KB 1.4.0
- QUALITY_LEVELS: 99
- TIME: Thu Dec 21 22:47:32 2017
- TRACE_ARRAY_MIN_INDEX: 0
- TRACE_ARRAY_MAX_INDEX: 6549
- TRIM: -1 -1 -1.000000e+000
- TRACE_PEAK_AREA_RATIO: -1.000000e+000
- CHEM: term
- DYE: big
- END_COMMENT
- BEGIN_DNA
- G 3 3
- C 2 18
- C 2 25
- C 5 34
- T 3 44
- C 5 57
- T 4 73
- G 9 82

16

pdf electroferograma

17

Seq, txt

P1AA_16s_16sF.txt

```
>171221-009_O14_P1AA_16s_16sF.ab1        543
GCCCTCTGCATCCCCCTGAGAGTTGGGCCTGCTCGGTGATTAATATTTAA
CAGCTGCGGTATTATAACTGTACTAAGGTAGCATAATAATTTGCTCTATA
AATTGGGGCTAGAATGAATGGTTTGACGAAAATTTGACTGTCTCTATTTT
ATTTATTAGAAATTAATTTTTGTAGTGAGAAAGCTTAAATTGTTTAAAGG
GACGAAAAGACCCTATTGAGCTTATATTATTTTATAAATTATATGATTGT
TGTTTATTTAGAATAATTTTGATTGGGGTGATCAAGGAATAAGATTGTTT
AAATTAACTTCCTTAATTAATTTGTTTTGTGGAAAAATAAACCAAATTTT
TTGCTTAGAAGATAAGTTACCATAGGGATAACAGCGTAATTTTTTTTGAG
AGTTCATATTGAAAAAAGAGATTGCGACCTCGATGTTGGATTAAAATTAA
CCTTAAGGTGAAGAGGCTTTATTAGGTAAATCTGTTCGATTTTTAAAATT
TTACGTGATCTGGTTCCAAAACCGGAAATATTATTTTTCCCTC
```

Archivos
Formato: fasta

```
>171221-009_O14_P1AA_16s_16sF.ab1        543
GCCCTCTGCATCCCCCTGAGAGTTGGGCCTGCTCGGTGATTAATATTTAA
CAGCTGCGGTATTATAACTGTACTAAGGTAGCATAATAATTTGCTCTATA
AATTGGGGCTAGAATGAATGGTTTGACGAAAATTTGACTGTCTCTATTTT
ATTTATTAGAAATTAATTTTTGTAGTGAGAAAGCTTAAATTGTTTAAAGG
GACGAAAAGACCCTATTGAGCTTATATTATTTTATAAATTATATGATTGT
TGTTTATTTAGAATAATTTTGATTGGGGTGATCAAGGAATAAGATTGTTT
AAATTAACTTCCTTAATTAATTTGTTTTGTGGAAAAATAAACCAAATTTT
TTGCTTAGAAGATAAGTTACCATAGGGATAACAGCGTAATTTTTTTTGAG
AGTTCATATTGAAAAAAGAGATTGCGACCTCGATGTTGGATTAAAATTAA
CCTTAAGGTGAAGAGGCTTTATTAGGTAAATCTGTTCGATTTTTAAAATT
TTACGTGATCTGGTTCCAAAACCGGAAATATTATTTTTCCCTC
```

18

FASTA

* Archivo de texto
* Secuencias de DNA
* Secuencias de amino ácidos (péptidos)
* Línea de descripción, que contiene identificadores
* Secuencia (una o varias líneas)

> Secuencia de amino ácidos

MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGK
AAKKADRLAAEGLVSVKVSDDFTIAAMRPSYLS

> Secuencia de DNA

ATGACTGAAATTACTGCTGCAATGGTAAAAGAACTCCGCGAAAGTACAG
GCGCGGGGATGATGGATTGTAAAAATGCTTTGAGTGAAACTAATGGAG
ATTTTGATAAAGCAGTACAACTTTTAAGAGAAAAAGGTTTAGGTAAGGC
TGCTAAAAAAGCAGATAGACTTGCTGCAGAAGGTTTGGTAAGTGTAAA
AGTA

19

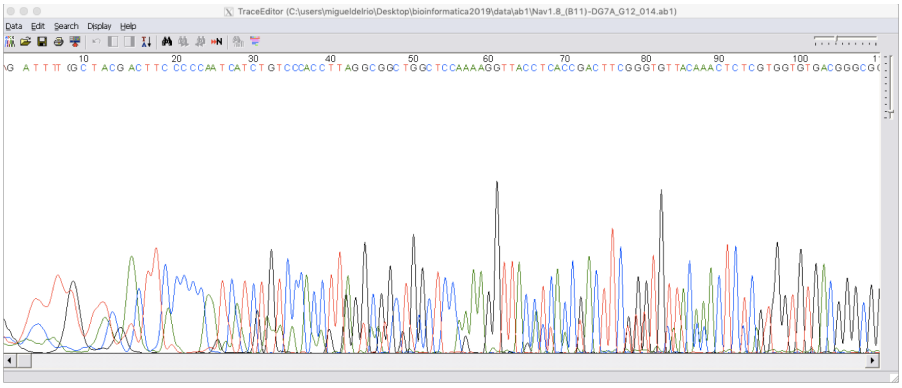
Secuencia de DNA (y RNA)

Código de ácido nucléico	Significado
A	Adenosina
C	Citosina
G	Guanina
T	Timidina
U	Uracilo
R	G A (puRina)
Y	T C (pirimidina/pYrimidine)
K	G T (cetona/Ketone)
M	A C (grupo aMino)
S	G C (interacción fuerte/**S**trong interaction)
W	A T (interacción débil/**W**eak interaction)
B	G T C (no A) (**B** viene tras la A)
D	G A T (no C) (**D** viene tras la C)
H	A C T (no G) (**H** viene tras la G)
V	G C A (no T, no U) (**V** viene tras la U)
N	A G C T (cualquiera/a**N**y)
X	máscara
-	hueco (gap) de longitud indeterminada

20

ab1

* Archivo con la información de salida (electroferograma) de la secuenciación Sanger del secuenciador de Applied Biosystems



csv

* Archivo de texto, valores separados por comas (comma separated values)

```
especie,org,long_total
especie1,1,100
especie1,2,89
especie1,3,114
especie1,4,88
especie1,5,124
especie1,12,101
```

fasta
(.fa, .fas, .fasta, .faa)

GenBank (nucleotido)

* >NM_001273529.2 Drosophila melanogaster vasa (vas), transcript variant B, mRNA
CACTAGATTTTTGGTACTTTTAACAGATCCTTTTCGGT
TTTGCGTTGCGCGAAGTGATCTGAACTTATCA
AAGTTTGTAAGGTAATACATAAAGTGAAAAAGAATTA
ATTTGCTCTTGAAAGGCAGGCCAAATTAAAAAA
AAAATATCAATATGTCTGACGACTGGGATGATGAGCC
CATTGTTGATACTCGCGGCGCCCGCGGTGGAGA
TTGGAGCGATGATGAGGACACGGCCAAGAGCTTCAG
CGGCGAAGCTGAAGGCGATGGTGTTGGAGGGAGC
GGTGGTGAAGGCGGCGGCTACCAAGGAGGAAATCGA
GATGTGTTCGGAAGGATCGGCGGAGGCAGAGGAG
GAGGAGCTGGAGGTTATCGAGGAGGAAATCGAGATG
GAGGGGGCTTCCACGGTGGACGTCGCGAGGGAGA
AAGGGACTTCCGCGGTGGAGAAGGCGGCTTCCGCGG
TGGACAAGGCGGCTCCCGCGGTGGACAAGGCGGC
TCCCGCGGGGGACAAGGCGGCTTCCGTGGTGGAGAA
GGCGGCTTCCGCGGTCGGCTGTACGAAAACGAGG

Swiss-Prot

* >tr|M9PBB5|M9PBB5_DROME Vasa, isoform B OS=Drosophila melanogaster GN=vas PE=3 SV=1
MSDDWDDEPIVDTRGARGGDWSDDEDTAKSFSGEAE
GDGVGGSGGEGGGYQGGNRDVFGR
IGGGRGGGAGGYRGGNRDGGGFHGGRREGERDFRGG
EGGFRGGQGGSRGGQGGSRGGQGG
FRGGEGGFRGRLYENEDGDERRGRLDREERGGERRG
RLDREERGGERGERGDGGFARRRR
NEDDINNNNNIVEDVERKREFYIPPEPSNDAIEIFSSGIA
SGIHFSKYNNIPVKVTGSDV
PQPIQHFTSADLRDIIIDNVNKSGYKIPTPIQKCSIPVISS
GRDLMACAQTGSGKTAAFL
LPILSKLLEDPHELELGRPQVVIVSPTRELAIQIFNEARK
FAFESYLKIGIVYGGTSFRH
QNECITRGCHVVIATPGRLLDFVDRTFITFEDTRFVVLD
EADRMLDMGFSEDMRRIMTHV

fastaq

* Archivos de texto con información de la secuenciación masiva
1. Encabezado "@",
2. secuencia,
3. Separador "+",
4. calidad de cada nucleótido

```
@M00313:49:000000000-A41BL:1:1101:15868:1583 1:N:0:356
GNATGAAATGTTCCAAAATGTCTTGGCATGCTGAAGCAGAGTTCCTTTCATCAGAACT
AACGGGCCCAAATCCTGAAAAACAACCCCACACCATAACCCCCCCACCACCACCACCA
AACTTTAAACTTGATACAGTGCAATCAGACAAATACTGTTCTCCTGGCAACTGCCAAA
CCCAGACTCATCCAGACAGAGAAGTGTGACTGGTCACTCCAGAGAATATGTGTAC
+
B#>>AABFFFDFGGGGGGGGGGHHHHAEGHFGGFHHGHHHHGHHHHHHHEGHFHHGHG
HHFEEGEEEEFGHFHHFGGHFC?FFF1EE?EF?1CGHGB3?>E/EEEGHHGHHGFHAG
HHGGHHHHHHFFDGGHHHHD<FDGFGHHHCHHFGHGFF1?=FFFFFGECH.<FFFCCH
H.<EGHEGCGC;GGHHECFHHHHGCG0;00;CE;BFFGGFFFGGGGFGBF;CFB0
```

gb

* Archivo con la información del GenBank

```
LOCUS       AB191108                503 bp    DNA     linear   INV 23-MAR-2005
DEFINITION  Argonauta argo gene for 16S rRNA, partial sequence.
ACCESSION   AB191108
VERSION     AB191108.1
KEYWORDS    .
SOURCE      Argonauta argo
  ORGANISM  Argonauta argo
            Eukaryota; Metazoa; Lophotrochozoa; Mollusca; Cephalopoda;
            Coleoidea; Neocoleoidea; Octopodiformes; Octopoda; Incirrata;
            Argonautidae; Argonauta.
REFERENCE   1
  AUTHORS   Takumiya,M., Kobayashi,M., Tsuneki,K. and Furuya,H.
  TITLE     Phylogenetic Relationships among coleoid cephalopods in Japanese
            waters
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 503)
  AUTHORS   Takumiya,M., Kobayashi,M., Tsuneki,K. and Furuya,H.
  TITLE     Direct Submission
  JOURNAL   Submitted (24-SEP-2004) Hidetaka Furuya, Osaka University,
            Department of Biology, Graduate School of Science; 1-1,
            Machikaneyama, Toyonaka, Osaka 560-0043, Japan
            (E-mail:hfuruya@bio.sci.osaka-u.ac.jp, Tel:81-6-6850-6775,
            Fax:81-6-6850-5817)
```

25

out

* Archivo de texto, usualmente se utiliza como indicador de salida (blast.out)
* No tiene un formato particular
* Adquiere el formato del programa que lo genera

26

tab

* Archivo separado por tabuladores. Similar a csv, pero en vez de comas se usa el tabulador

Phred score
Es una medida de calidad en la identificación de los nucleótidos generados por algún método de secuenciación.

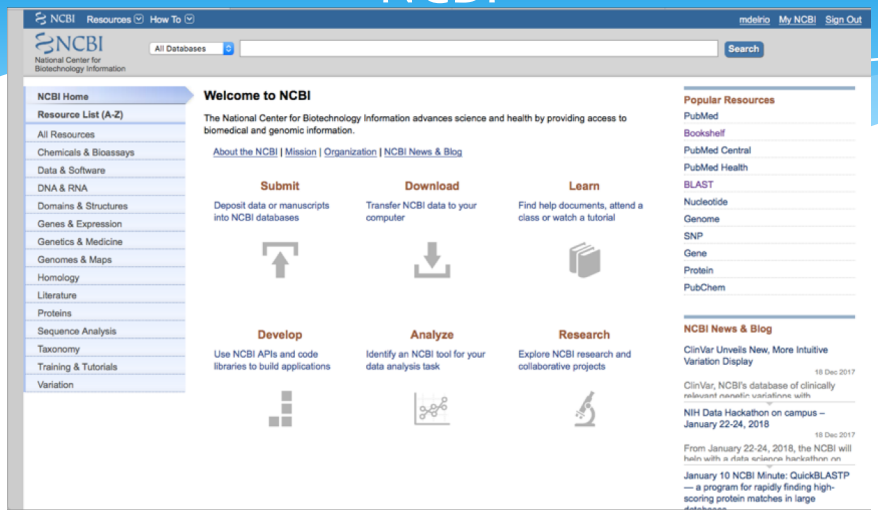
El *Phred score* o *Q* se define como una propiedad que está relacionada logarítmicamente con las probabilidades de error en la identificación de las bases (*P*).

$$Q = -10 \log_{10} P$$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

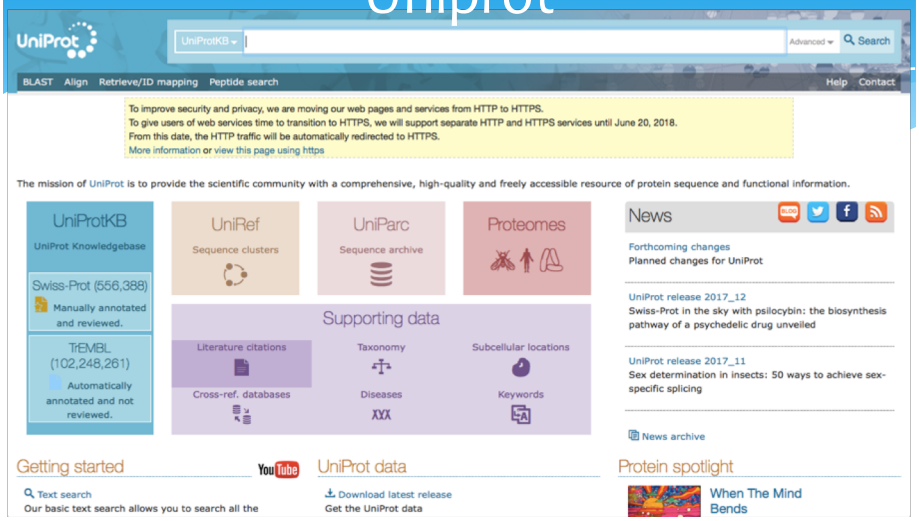
Sanger

NCBI



29

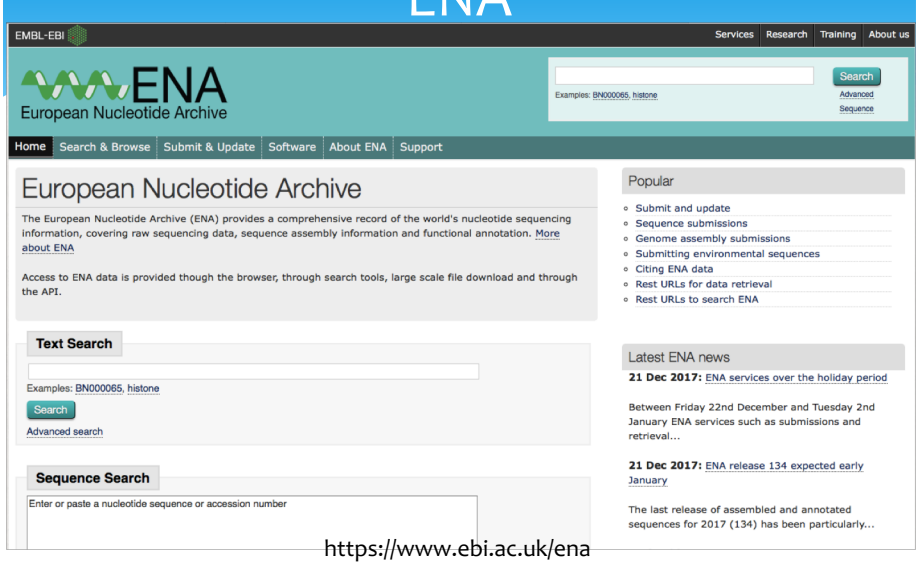
Uniprot



http://www.uniprot.org

30

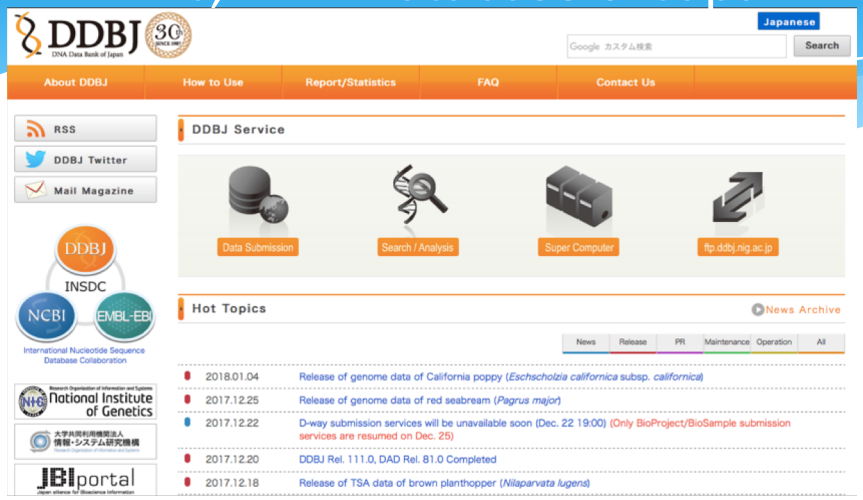
ENA



https://www.ebi.ac.uk/ena

31

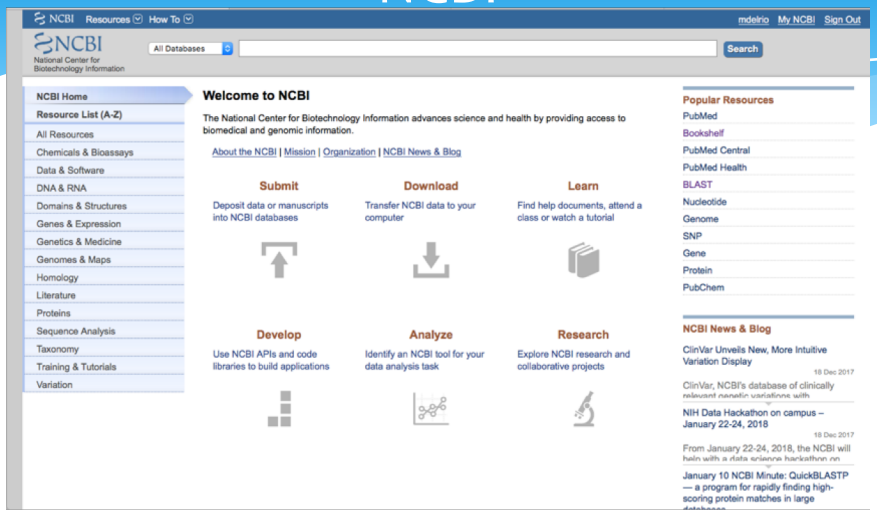
DDBJ, DNA Database of Japan



http://www.ddbj.nig.ac.jp

32

NCBI



33

Bases de datos

Nombre	Dirección	Datos
NCBI	https://www.ncbi.nlm.nih.gov	
GenBank	https://www.ncbi.nlm.nih.gov/genbank/	Base de datos
Nucleotide	https://www.ncbi.nlm.nih.gov/nuccore/	nucleótidos
PubMed	https://www.ncbi.nlm.nih.gov/pubmed	Referencias bibliográficas
NR	https://www.ncbi.nlm.nih.gov/protein/	Secuencia de proteínas no redundante
Swiss-Prot, UniProtKB	http://www.uniprot.org	Universal Protein Resource
KEGG (Kyoto Encyclopedia of Genes and Genomes)	http://www.genome.jp/kegg/	Rutas metabólicas

34

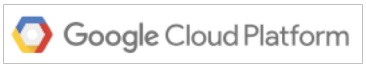
Identificadores en diferentes bases de datos

Database name	Identifier syntax
GenBank	accession|locus (gb|accession|locus)
EMBL Data Library	emb|accession|locus
DDBJ, DNA Database of Japan	dbj|accession|locus
NBRF PIR	pir||entry
Protein Research Foundation	prf||name
SWISS-PROT	sp|accession|entry name
Brookhaven Protein Data Bank	pdb|entry|chain
Patents	pat|country|number
GenInfo Backbone Id	bbs|number
General database identi era	gnl|database|identi er
NCBI Reference Sequence	ref|accession|locus
Local Sequence identi er	lcl|identifier

McEntyre, Jo, Ostell. 2012. The NCBI Handbook
https://www.ncbi.nlm.nih.gov/books/NBK21101/pdf/Bookshelf_NBK21101.pdf

Datos en servidores (nube)

* Servicios de almacenamiento
* Servicios de renta de equipo
* Servicio de procesamiento



37

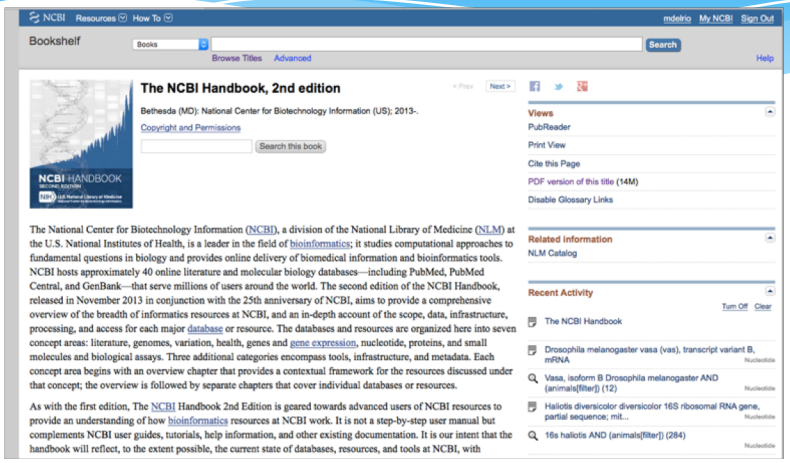
Recomendaciones al usar servidores

1. Los datos nunca están seguros en la red
2. Anotar el servidor, la base de datos y la versión del programa utilizado
3. Anotar los números de identificación de las secuencias
4. Anotar los parámetros del programa
5. Guarde los resultados de internet inmediatamente
6. Utilice los valores de E
7. Asegúrese de que puede confiar en los alineamientos
8. Utilice diferentes programas para verificar los resultados extremos
9. No utilice métodos que no se han publicado
10. Las bases de datos no son como los vinos (actualícela)
11. No confíe ciegamente en los programas gratuitos
12. Utilice las herramientas en el momento adecuado (Bitting the bullet at the right time)

Claverie y Notredame. 2007. Bioinformatics for dummies. 2nd Ed. Wile Publishing, Inc.

38

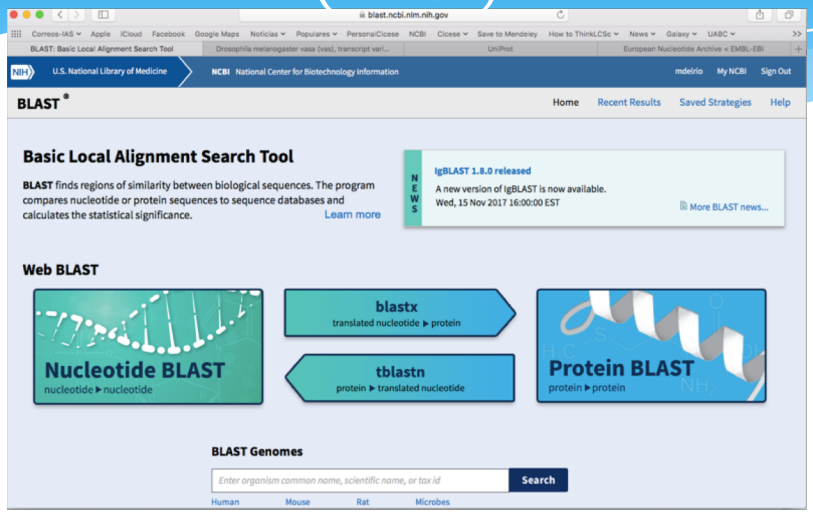
Manuales



https://www.ncbi.nlm.nih.gov/books/NBK143764/

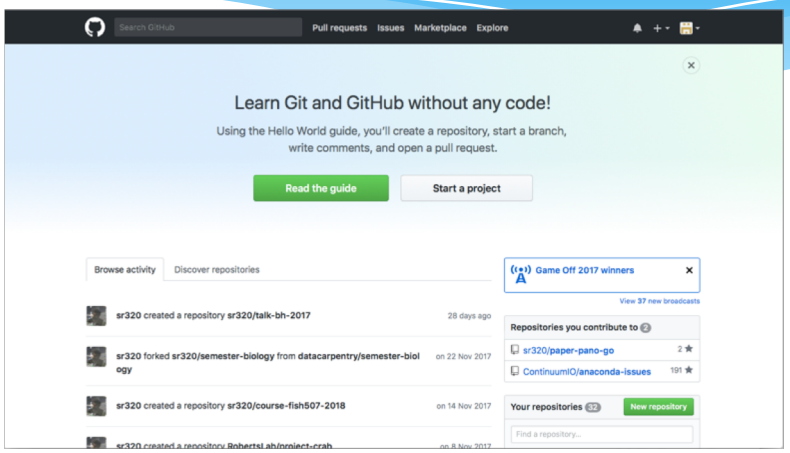
39

Basic Local Alignment Search Tool (BLAST)



40

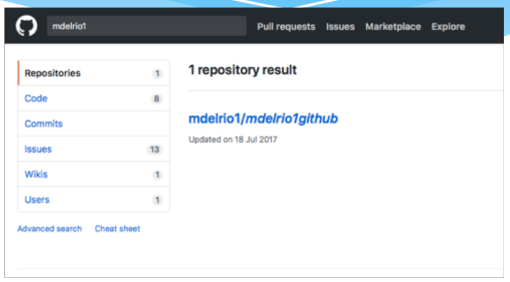
Github



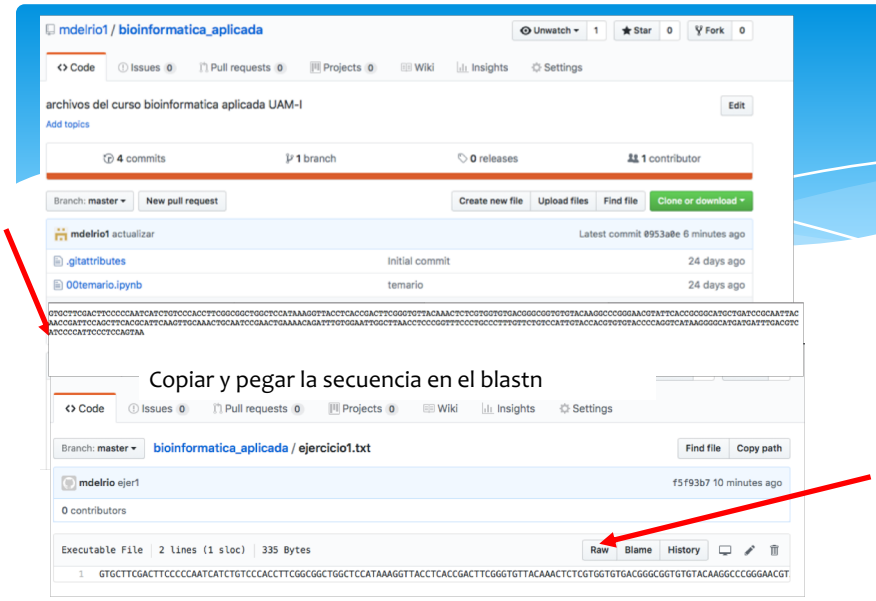
41

Github

* Abrir una cuenta
* Buscar mdelrio1
* En usuarios, seguirme

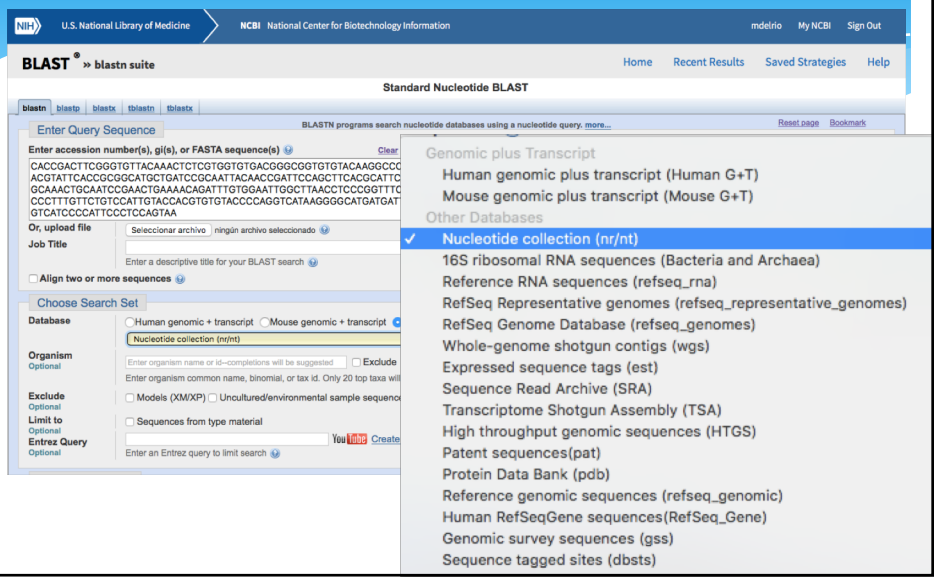


42


mdelrio1 / bioinformatica_aplicada
archivos del curso bioinformatica aplicada UAM-I
Add topics
4 commits
1 branch
0 releases
1 contributor
Clone or download
mdelrio1 actualizar
.gitattributes
Initial commit
24 days ago
00temario.ipynb
temario
24 days ago
Copiar y pegar la secuencia en el blastn
bioinformatica_aplicada / ejercicio1.txt
Find file
Copy path
mdelrio ejer1
0 contributors
Executable File
2 lines (1 sloc)
335 Bytes
Raw
Blame
History

43

Blastn


U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
BLAST » blastn suite
Home
Recent Results
Saved Strategies
Help
Standard Nucleotide BLAST
Enter Query Sequence
Choose Search Set
Database
Organism
Exclude
Limit to
Entrez Query
Genomic plus Transcript
Human genomic plus transcript (Human G+T)
Mouse genomic plus transcript (Mouse G+T)
Other Databases
Nucleotide collection (nr/nt)
16S ribosomal RNA sequences (Bacteria and Archaea)
Reference RNA sequences (refseq_rna)
RefSeq Representative genomes (refseq_representative_genomes)
RefSeq Genome Database (refseq_genomes)
Whole-genome shotgun contigs (wgs)
Expressed sequence tags (est)
Sequence Read Archive (SRA)
Transcriptome Shotgun Assembly (TSA)
High throughput genomic sequences (HTGS)
Patent sequences(pat)
Protein Data Bank (pdb)
Reference genomic sequences (refseq_genomic)
Human RefSeqGene sequences(RefSeq_Gene)
Genomic survey sequences (gss)
Sequence tagged sites (dbsts)

Ejercicio 1

* Someta la secuencia al GenBank

CTTCCCCCAATCATCTGTCCCACCTTCGGCGGCTGGCTCCATAAA
GGTTACCTCACCGACTTCGGGTGTTACAAACTCTCGTGGTGTGAC
GGGCGGTGTGTACAAGGCCCGGGAACGTATTCACCGCGGCATGC
TGATCCGCAATTACAACCGATTCCAGCTTCACGCATTCAAGTTGC
AAACTGCAATCCGAACTGAAAACAGATTTGTGGAATTGGCTTAA
CCTCCCGGTTTCCCTGCCCTTTGTTCTGTCCATTGTACCACGTGTG
TACCCCAGGTCATAAGGGGCATGATGATTTGACGTCATC

* Utilizando las bases de datos
  * Nt
  * 16S microbial
* Con sus compañeros de mesa discuta las diferencias entre la información proporcionada en cada búsqueda

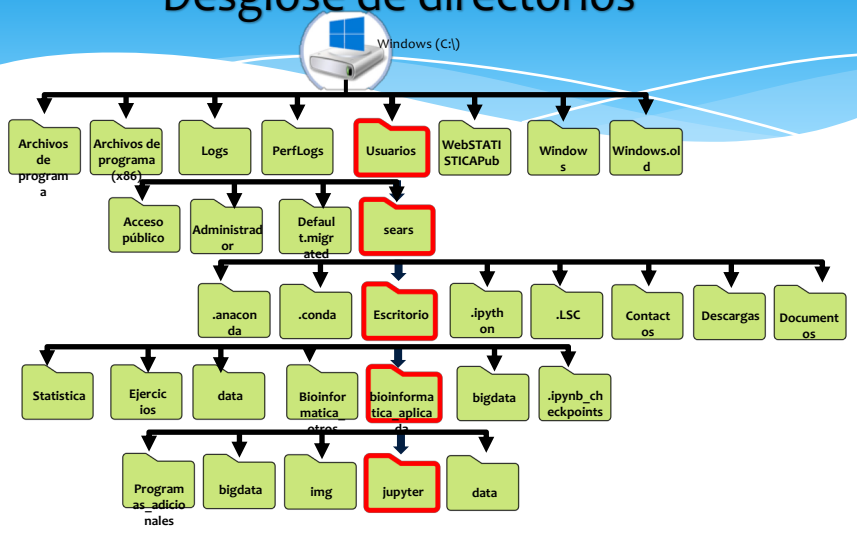
45

Cuestionario

* Defina los siguientes términos:
  * Alineamiento
  * Marcador/Calificación máxima (max score)
  * Marcador total (total score)
  * Cobertura de búsqueda
  * Valor de E (E value)
  * Identidad (ident)
  * Acceso
* ¿En qué formatos de archivo se puede descargar la información?

46

Desglose de directorios

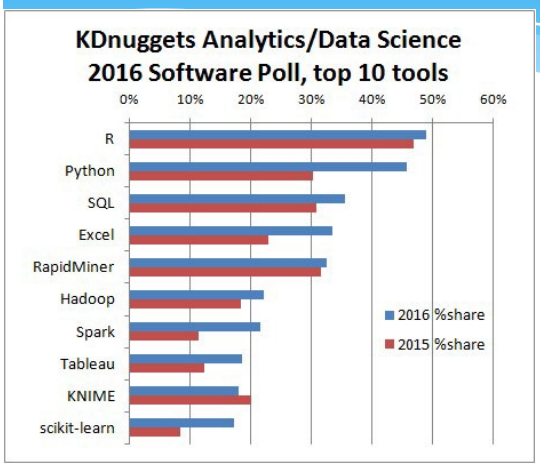


Lenguajes

* Python
* R
* Perl
* Ruby
* Java
* Linux
* C++

48

Python



Python es el lenguaje más popular utilizado por "científicos de datos"
Python está ganando popularidad en apreciación general y es el lenguaje más práctico para construir productos
Python es una herramienta poderosa para procesamiento de datos a mediana escala.
Python también tiene la ventaja de una comunidad rica y que ofrece una gran cantidad de herramientas y kits
Python no es el lenguaje con mayor

https://dzone.com/articles/which-are-the-popular-languages-for-data-science

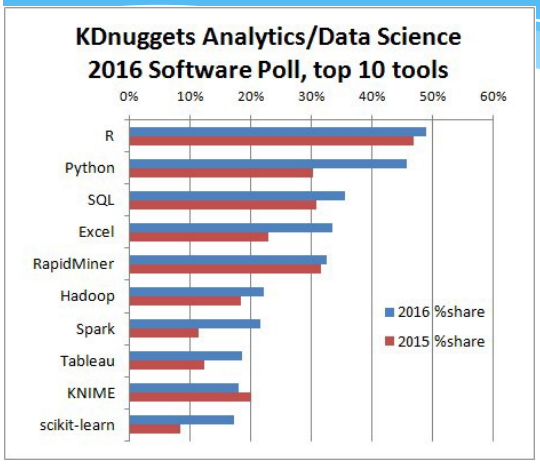
- Python
  - python.org
  - http://www.python-course.eu/index.php
- Google
  - https://developers.google.com/edu/python/

* Python tiene una comunidad sana, activa y de gran soporte
* Python tiene corporaciones que dan gran soporte
* Python tiene "Big Data"
* Python tiene un gran número y sorprendente de biblioteca (libraries).
* Python es confiable y eficiente
* Python es accesible

http://www.kdnuggets.com/2017/07/6-reasons-python-suddenly-super-popular.html

51

R



R ha estado desde 1997 como una alternativa gratuita de programas estadísticos caros, como Matlab o SAS. En los últimos años, R se ha vuelto el niño dorado de la ciencias de datos (Data Science)
Lo usan Google, Facebook, Bank of America, y the *New York Times*. Sus utilerías comerciales se han dispersado enormemente.

52

https://dzone.com/articles/which-are-the-popular-languages-for-data-science

Otros lenguajes

* Perl
* Ruby

53

Perl vs. Python

Use Python

* Cuando necesite usar el código más de una vez
* Siempre que haya una posibilidad remota que alguien más (i.e. colegas) utilicen su código
* Siempre que necesite usar funciones u objetos. La mayoría de los códigos de Perl no necesariamente incluyen las funciones, porque son más difíciles de escribir que en Python. Por ejemplo en Perl, necesita aprender cómo pasar de referencias a variables, etc. y ello conduce a códigos que son más difíciles de ententer (en Perl).
* Si es el primer lenguaje de programación que está aprendiendo. Se sugiere que inicie con Python,
* Python es mucho más limpio que Perl
* Python está diseñado para respetar buenas prácticas que cualquier programador debería conocer (ver The Zen of Python, https://www.python.org/dev/peps/pep-0020/)
* Python tiene soporte para estructura tabular de datos (Data Frames), siempre que tenga trabajos con tablas o análisis de datos use Python (Pandas) o R
* Python tiene soporte para "machine learning algorithms".

54

Giovanni M Dall'Olio https://www.biostars.org/p/13972/

Python

* "Es uno de los lenguajes de programación más populares hoy en día a pesar de ser un idioma relativamente viejo. Fue creado a finales de los ochentas por Guido van Rossum --que dentro de la comunidad de Python es conocido como el Benevolente Dictador Vitalicio-- y su nombre está inspirado en el grupo de comedia británico Monthy Python."

https://hipertextual.com/2011/02/zen-python

Python vs Ruby vs Perl

* Perl for one-liners in a pipe. Python for everything else
* Python vs Ruby vs Perl
* Right folks! Time for another religious war. Or, a Halo style three way battle.
* On the left, we have Perl, the grand-daddy of the scripting world. But don't let his age fool you, he still has a few tricks up his sleeve.
* On the right we have Ruby, the young kid on the block. He may be the youngest, but he sure knows how to throw his weight around.
* In the middle we have Python. Entering middle age, his pot belly is beginning to show, but get too close, and he'll knock you out in one.
* Come ladies and gentlemen, place bets. Who will win?

56

El zen de Python:

Hermoso es mejor que feo.
Explícito es mejor que implícito.
Simple es mejor que complejo.
Complejo es mejor que complicado.
Sencillo es mejor que anidado.
Escaso es mejor que denso.
La legibilidad cuenta.
Los casos especiales no son lo suficientemente especiales para romper las reglas.
Lo práctico le gana a la pureza.
Los errores no debe pasar en silencio.
A menos que sean silenciados.
En cara a la ambigüedad, rechazar la tentación de adivinar.
Debe haber una - y preferiblemente sólo una - manera obvia de hacerlo.
Aunque esa manera puede no ser obvia en un primer momento a menos que seas holandés.
Ahora es mejor que nunca.
Aunque "nunca" es a menudo mejor que "ahora mismo".
Si la aplicación es difícil de explicar, es una mala idea.
Si la aplicación es fácil de explicar, puede ser una buena idea.
Los espacios de nombres son una gran idea ¡hay que hacer más de eso!

Hermoso es mejor que feo.
Explícito es mejor que implícito.
Simple es mejor que complejo.
Complejo es mejor que complicado.
Plano es mejor que anidado.
Escaso es mejor que denso.
Cuenta la legibilidad.
Los casos especiales no son lo suficientemente especial como para romper las reglas.
Aunque sentido práctico supera pureza.
Los errores nunca debe pasar en silencio.
A menos que explícitamente silenciados.
Ante la ambigüedad, rechaza la tentación de adivinar.
Debería haber una - y preferiblemente sólo una - manera obvia de hacerlo.
Aunque esa manera puede no ser obvia al principio a menos que seas holandés.
Ahora es mejor que nunca.
Aunque nunca es a menudo mejor que la * justo * ahora.
Si la implementación es difícil de explicar, es una mala idea.
Si la implementación es fácil de explicar, puede ser una buena idea.
Namespaces son una gran idea de fanfarria - Vamos a hacer más de esos!

57

Jupyter

https://jupyter.readthedocs.io/en/latest/install.html#install

58

Anaconda

https://www.continuum.io/downloads