

Procesado y Manejo de Datos Masivos

Resumen

En esta asignatura se aprenderán técnicas para procesar y manipular grandes volúmenes de datos utilizando programación y línea de comandos. Se abordará el manejo y análisis de formatos de datos comunes en bioinformática, como FASTA, GFF y GenBank, utilizando patrones que optimizan el uso de memoria y tiempo de ejecución, así como técnicas de programación paralela. Además, se realizarán operaciones avanzadas en bases de datos relacionales y no relacionales, y se explorará el acceso programático a bases de datos biomédicas online a través de sus APIs.

Índice general

I	Estrategias de programación para el parseo y extracción de datos	3
I.1	Formatos de ficheros bioinformáticos	3
I.1.1	FASTA	3
I.1.2	FASTQ	5
I.1.3	GFF: generic feature format	6
I.1.4	GenBank record	9
I.2	Librerías de Python	9
I.2.1	NumPy	9
I.2.2	Matplotlib	16
I.3	Biopython	20
I.3.1	Seq	20
I.3.2	SeqRecord	21
I.3.3	SeqIO	22
I.3.4	Atributos de SeqRecord	23
II	Acceso programático a bases de datos biomédicas	26
II.1	Introducción	26
II.1.1	Acceso programático en formularios	27
II.2	JSON	28
II.3	Protocolo HTTP	29
II.3.1	Aspectos básicos de HTTP	29
II.3.2	Flujo HTTP	29
II.3.3	APIs con HTTP	31
II.4	API REST	32
II.5	Examen de prueba	33
II.5.1	Ejercicio 1	33
II.5.2	Ejercicio 2	34
III	La línea de comandos de Linux	36
III.1	Introducción	36
III.1.1	Test inicial	36
III.1.2	Redirección y tuberías	37
III.1.3	Filtros	37
III.1.4	Scripts de shell	39
III.1.5	Editores de texto	40
III.2	Expresiones regulares y grep	41
III.2.1	Expresiones reguladas extendidas	42
III.3	Expansiones	43
III.3.1	Expansión aritmética	43

III.3.2 Expansión de llaves	43
III.4 Sustitución de comandos	43
III.5 El comando sed	44
III.6 El comando awk	44
IV Bases de datos no relacionales	47
IV.1 Bases de datos no SQL	47
IV.1.1 Motivación para las bases de datos no SQL	47
IV.1.2 Not only SQL	48
IV.2 Pajek	51

Capítulo I

Estrategias de programación para el parseo y extracción de datos

I.1. Formatos de ficheros bioinformáticos

Los formatos de fichero son en formato texto, y los principales son:

- FASTA: guarda secuencia de nucleótidos o aminoácidos.
- FASTQ: conjunto de secuencias de nucleótidos leídas junto con sus puntuaciones de calidad
- GFF/GTF: formato de anotación de características del genoma
- GenBank: formato enriquecido para secuencias de nucleótidos o aminoácidos.

I.1.1. FASTA

Guarda una o varias secuencias de nucleótidos o aminoácidos. Se pronuncia como "fast A", que significa "fast all", siendo all cualquier alfabeto. Los archivos fasta pueden tener las siguientes extensiones:

- **.fasta, .fa**: extensión FASTA genérica
- **.fna**: contiene secuencia de nucleótidos
- **.faa**: contiene secuencia de aminoácidos
- **.frn**: contiene secuencias de ARN no codificante

En la figura [I.1](#) se puede ver la estructura de un fichero FASTA. Cada secuencia consiste en una línea descriptiva seguida de las líneas que contienen la secuencia. La línea descriptiva cuenta con un símbolo > seguido del identificativo de la secuencia. Tras un espacio, se pone la descripción de la secuencia. En algunos casos, las secuencias se encuentran en mayúscula, en minúscula o en una mezcla de ambas. Por ello, se

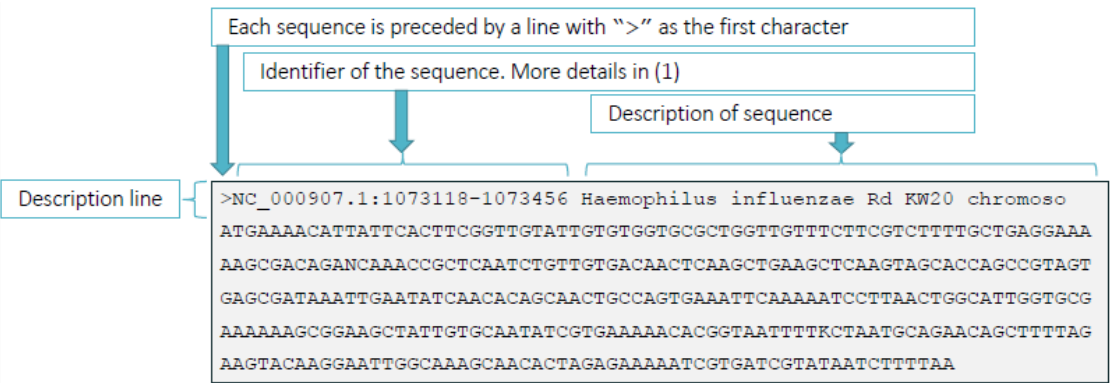


Figura I.1: Anatomía de un fichero FASTA

puede pasar toda la secuencia a mayúsculas o minúsculas con `.upper()` y `.lower()` respectivamente.

La secuencia se puede expandir por múltiples líneas, las cuales generalmente tienen unos 80 caracteres. No hay espacios en las líneas o líneas en blanco en una secuencia. Cada elemento de una secuencia se describe con un único carácter, y normalmente se escriben en mayúscula (aunque esto es una convención, no es estándar). No es recomendable utilizar guiones para representar gaps, ya que hay algunos programas que no los pueden manejar. Es mejor utilizar una N en el caso de secuencias de nucleótidos y X de aminoácidos (véase tabla I.1).

FASTA for genome data		
A adenosine	C cytidine	G guanine
T thymidine	N A/G/C/T (any)	U uridine
K G/T (keto)	S G/C (strong)	Y T/C (pyrimidine)
M A/C (amino)	W A/T (weak)	R G/A (purine)
B G/T/C	D G/A/T	H A/C/T
V G/C/A	- gap	

FASTA for protein data		
A alanine	B aspartate/asparagine	C cystidine
D aspartate	E glutamate	F phenylalanine
G glycine	H histidine	I isoleucine
K lysine	L leucine	M methionine
N asparagine	P proline	Q glutamine
R arginine	S serine	T threonine
U selenocysteine	V valine	W tryptophan
Y tyrosine	Z glutamate/glutamine	X any
* translation stop	- gap	

Tabla I.1: Caracteres en los ficheros FASTA para datos genómicos y de proteínas.

I.1.1.1. Parsear ficheros FASTA en Python

En informática, parsear significa leer un fichero. Se pueden leer ficheros FASTA en Python de la siguiente forma:

```
def readFasta(file):
    """ Reads all sequences of a FASTA file
        returns a dictionary """
    d = {}
    identifier = ''

    with open(file, 'r') as f:
        for linea in f:
            if linea[0] == ">": #alternativa: linea.startswith(">")
                if identifier != '':
                    d[identifier] = ''.join(d[identifier])
                    descriptors = linea[1:].split()
                    identifier = descriptors[0]
                    #Alternativa: identifier = linea[1:linea.find(' ')]
                    d[identifier] = []

            else:
                d[identifier].append(linea.strip('\n'))
        d[identifier] = ''.join(d[identifier])
    return d

readFasta("phix174/phix.fa")
```

I.1.2. FASTQ

Los ficheros FASTQ guardan secuencias de nucleótidos junto con las calidades.

Cada secuencia de un archivo FASTQ consta de 4 líneas. La primera es la línea de descripción precedida de . Tiene un formato libre sin límite de longitud. A veces se coloca un identificador justo después de la . La segunda línea es la línea de secuencia, e incluye la secuencia propiamente dicha. Sigue normas y convenciones similares a las del fichero FASTA: normalmente en mayúsculas, sin espacios permitidos, etc. Esta línea podría estar envuelta en múltiples líneas como en un archivo FASTA. La siguiente es la línea que señala el final de la secuencia. Está marcada con el signo +. Puede ir seguida de la misma descripción dada en la primera línea. Sin embargo, es más sensato dejarla sólo con el '+' para reducir el tamaño de los ficheros. La última línea es la línea de puntuaciones de calidad: una secuencia codificada en caracteres con la calidad de las lecturas de la secuencia. Contiene un carácter por base. Los caracteres permitidos son como máximo ASCII 33-126 inclusive, que son todos imprimibles. Esta línea puede ser envuelta como en un archivo FASTA y su longitud total debe ser igual a la secuencia. Véase un ejemplo de un archivo FASTQ en la figura [I.2](#).

Hay varias cuestiones que deben tenerse en cuenta al analizar un archivo FASTQ: Tanto el carácter como + pueden aparecer en la línea con los índices de calidad. Pueden aparecer en la primera posición, por lo que hay que tener cuidado al utilizar

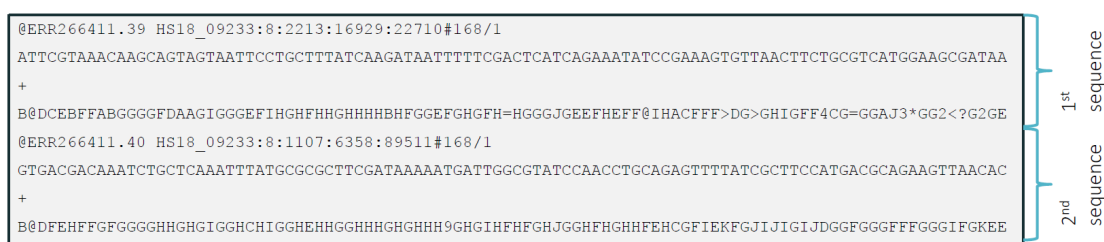


Figura I.2: *Anatomía de un fichero FASTQ*

grep o herramientas similares: por ejemplo `$> grep -v "[+]" file.fastq` (Este comando para eliminar las líneas `+` y no es una buena idea). En principio, tanto la secuencia como las puntuaciones de calidad pueden tener múltiples saltos de línea. El programa de análisis debería eliminar los saltos de línea. Para reducir los problemas de análisis sintáctico, la mayoría de las herramientas producen ahora archivos fastq en los que las secuencias y las puntuaciones de calidad se dan en una sola línea, posiblemente muy larga. Por lo tanto, cada secuencia tiene exactamente 4 líneas, lo cual es conveniente.

La codificación más utilizada para las puntuaciones de calidad es el formato Sanger o codificación Phred+33. Cada carácter codifica la calidad de la lectura, Q, que se define como:

$$Q = -10 \cdot \log_{10} p$$

siendo p la probabilidad de error de la lectura. Esta puntuación de calidad se utiliza porque permite una interpretación fácil de la probabilidad ($Q = 10$, $p = 1/10$; $Q = 20$, $p = 1/100$; $Q = 30$, $p = 1/1000$) y porque permite una codificación en un único carácter ASCII. No obstante, no se puede guardar Q directamente como un carácter porque no todos son imprimibles. Por ello, se emplea la codificación Phred+33, que codifica el valor Q como:

$$Q = ord(chr) - 33$$

donde `ord()` convierte el carácter a su representación ASCII. Por ejemplo, el símbolo tiene el código ASCII de 64, por lo que su `Q` sería de 31. Una `p` de 0 significa que es una lectura segura, mientras que una `p` de 1 indica que la lectura no es nada segura. Para calcular la `p`, se despeja la fórmula para que quede:

$$p = 10^{-\frac{Q}{10}}$$

Ejemplos: Para el carácter " ", el código ASCII es 126, por lo que su Q es $126 - 33 = 93$. En este caso, la p sería de $0,5 \cdot 10^{-10}$. Así, el carácter "!", cuyo ASCII es 33, tiene una Q de 0 y una p de 1.

En este caso, como las letras en mayúscula y minúscula tienen un código ASCII diferente, no se puede pasar todo a mayúsculas o minúsculas como en FASTA.

I.1.3. GFF: generic feature format

Los ficheros GFF tienen un formato de texto sin formato para representar características genómicas. Cada característica está representada por nueve columnas separadas por tabuladores:

1. seqid: id de la secuencia donde se localiza la característica. No se permiten espacios
2. source: nombre del programa o algoritmo que ha generado esta característica.
3. type: tipo de característica (region, gene, CDS, etc.)
4. start: posición donde inicia la característica, su primera base
5. end: posición donde finaliza la característica, su última base
6. score: un valor flotante que indica la confianza en la característica. Si no se conoce, se pone un punto.
7. strand: indica el sentido de la cadena, siendo + la cadena positiva y - la cadena negativa.
8. phase: puede ser 0, 1 o 2 para CDS y un punto para todo lo demás.
9. attributes: Una lista de atributos de la característica en formato tag=value separados por punto y coma (;). Se permiten espacios en blanco en estas columnas, así como tabuladores y punto y coma si se escapan. Los atributos son:
 - **ID**: El id de la característica. Debe ser único dentro del archivo gfffile. Tenga en cuenta que para las características no contiguas el id puede aparecer en varias líneas. Se considera una característica única
 - **Name**: Nombre de la característica. No es necesario que sea único.
 - **Parent**: Id del elemento padre. Indica la relación «parte de». Un elemento puede tener varios padres. En este caso, los identificadores se separan por comas.
 - **Note**: nota de texto libre
 - ...

```
##gff-version 3
##...
NC_000907.1 RefSeq region 1 1830138 . + . ID=id0;Dbxref=taxon:71421;Is_circular=true;Name=...
NC_000907.1 RefSeq remark 1 1830138 . + . ID=id1;Note=REFSEQ gene predictions performed by...
NC_000907.1 RefSeq gene 2 1021 . + . ID=gene0;Dbxref=GeneID:950899;Name=gapdH;gbkey=G...
NC_000907.1 RefSeq CDS 2 1021 . + 0 ID=cds0;Parent=gene0;Dbxref=Genbank:NP_438174.1,...
```

} First four features

Figura I.3: *Primeras columnas de un fichero GFF.*

Hay varias cuestiones que deben tenerse en cuenta al analizar un archivo GFF: Las columnas sólo están separadas por tabuladores, pero se permiten espacios en blanco en algunos campos. Por tanto, hay que tener cuidado con grep y herramientas shell similares. Algunos caracteres se escapan en algunos campos, es decir, no queremos que los caracteres se interpreten como los caracteres que son. Esto ocurre con los puntos y coma y los espacios. Estos caracteres se escriben de otra forma; por ejemplo, un valor

de a en punto y coma debe codificarse con %3B. Para decodificar estos caracteres puede utilizar urllib. Además, una fila con triple almohadilla (###) indica que se han resuelto todas las características multilinea hasta ese punto. Hay al menos una de estas líneas al final del fichero. Las líneas que empiezan por # son comentarios. La región definida por los rasgos CDS incluye los codones de inicio y fin.

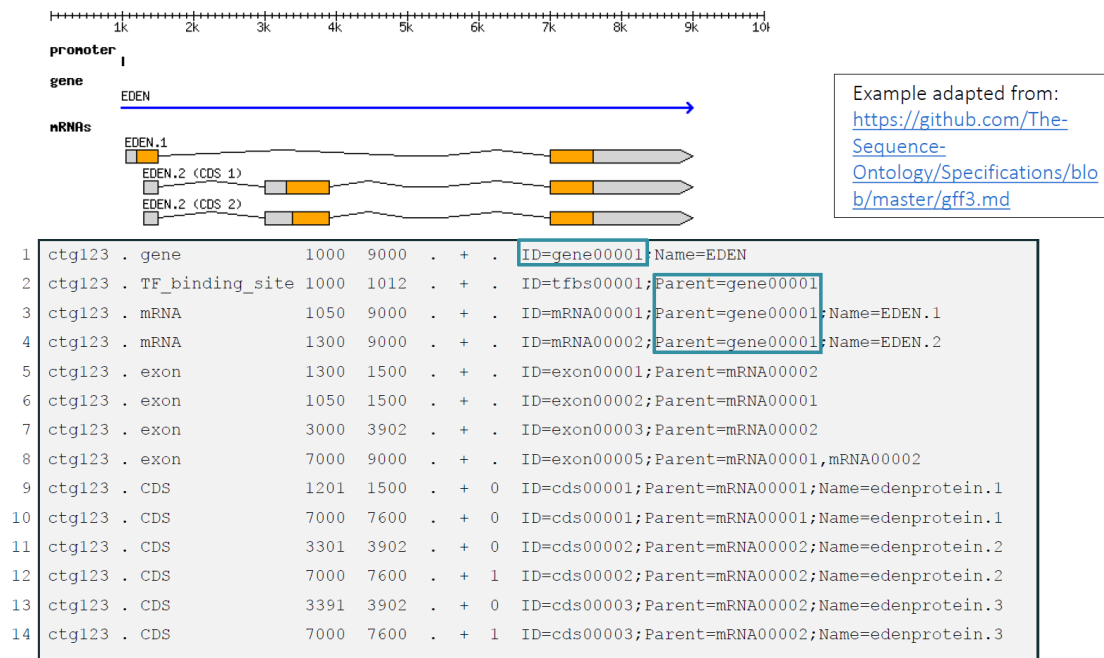


Figura I.4: Ejemplo gráfico de una región genómica con el fichero GFF.

Se puede trabajar con un fichero GFF desde la terminal de Linux. Por ejemplo, se pueden buscar todas las regiones (`grep region fichero.gff`) u obtener las terceras columnas (`cut -f3 fichero.gff`).

```
def readGFF(file):
    """ Counts region types of a GFF file
        returns a dictionary """
    d = {}

    with open(file, 'r') as f:
        for linea in f:
            if linea[0] != "#":
                campos = linea.split("\t")
                region_type = campos[2]
                if region_type not in d: #o d.keys()
                    d[region_type] = 1
                else:
                    d[region_type] += 1
    return d

readGFF('data/haemophilus_influenzae/GCF_000027305.1_ASM2730v1_genomic.gff')
```

I.1.4. GenBank record

GenBank es una base de datos de secuencias de libre acceso mantenida por el National Center for Biotechnology Information (NCBI). Esta base de datos tiene un formato de archivo plano para representar un registro en la base de datos. Este formato codifica la secuencia, las características y las referencias de proteínas y genes.

LOCUS	NC_000907	1656 bp	DNA	linear	CON 02-AUG-2016
DEFINITION	Haemophilus influenzae Rd KW20 chromosome, complete genome.				
ACCESSION	NC_000907 REGION: complement(36376..38031)				
VERSION	NC_000907.1				
DBLINK	BioProject: PRJNA57771				
	Assembly: GCF_000027305.1				
KEYWORDS	RefSeq.				
SOURCE	Haemophilus influenzae Rd KW20				
ORGANISM	Haemophilus influenzae Rd KW20				
	Bacteria; Proteobacteria; Gammaproteobacteria; Pasteurellales;				
	Pasteurellaceae; Haemophilus.				

Figura I.5: Principio de un documento GenBank.

Las partes de un documento GenBank (figura I.5) son:

- **LOCUS:** Contiene un identificador (NC_000907 en el ejemplo siguiente), la longitud de la secuencia (1656 pb), el tipo de molécula (ADN) y la fecha de la última modificación del registro.
- **DEFINITION:** Breve descripción de la secuencia.
- **ORGANISM:** Nombre científico del organismo .
- **REFERENCE:** Publicación científica relacionada con la secuencia. La última suele referirse al autor que envió la secuencia (esta referencia se marca con «direct submission» en lugar del título del artículo).
- **FEATURES:** Información sobre genes y productos génicos de la secuencia. Tiene varios subcampos:
 - **Source:** Campo obligatorio que describe la longitud y el organismo de la secuencia.
 - **gene:** región de un gen
 - **CDS:** región codificante de una proteína. Incluye translation, tabla, protein_id, productname, notas, etc.

I.2. Librerías de Python

I.2.1. NumPy

NumPy (Numerical Python) es una librería de Python que permite realizar operaciones con arrays y matrices. También incluye herramientas de álgebra lineal, estadística o generación aleatoria de números. Se combina frecuentemente con otras librerías como SciPy para computaciones numéricas y pandas para análisis de datos.

1.2.1.1. Clase ndarray

Creación de arrays El elemento principal de NumPy es el objeto **ndarray** que representa arrays multidimensionales. Es capaz de realizar operaciones vectoriales eficientes (en velocidad y memoria). Al contrario que las listas de Python, todos los elementos de un ndarray son del mismo tipo. Los tipos posibles son int, float, boolean, string, etc. El atributo shape te da el tamaño del array y dtype el tipo de los elementos.

```
import numpy as np
```

```
data = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
```

```
data.shape
```

```
#Output: (3, 3)
```

```
data.dtype
```

```
#Output: dtype('int64')
```

Se puede crear un array a partir de cualquier otra secuencia (listas u otros ndarrays), utilizando la función `numpy.array`. Es posible especificar en la construcción el tipo del array (dtype) y la dimensión. Las dimensiones son importantes para cuando haya que fusionar arrays, ya que deberán coincidir. Además, la dimensión debe coincidir en todo el array, es decir, debe ser homogéneo.

```
x = np.array([1, 2, 3], ndmin = 2)
```

```
# array([[1, 2, 3]])
```

```
y = np.array([1, 2], dtype = 'int32')
```

```
error = np.array([1, 2, 3], [4, 5])
```

Otras funciones para crear e inicializar arrays/matrices son:

- Matrices de ceros: `numpy.zeros` y `numpy.zeros_like` (el primero crea un array de las dimensiones dadas, el segundo crea un array como el que se le pasa, es decir, con sus dimensiones y el tipo).
- Matrices de unos: `numpy.ones` y `numpy.ones_like`
- Matriz identidad: `numpy.eye` y `numpy.identity`
- La función `numpy.arange` se comporta de forma similar a la función estándar de python `range`, pero devuelve un ndarray.
- La función `numpy.linspace` devuelve números espaciados uniformemente en un intervalo especificado. Se proporciona el primer y último valor y el número total de valores. La función `numpy.logspace` hace algo similar pero de forma logarítmica.

```
np.zeros((3,2))
```

```
#array([[0., 0.], [0., 0.], [0., 0.]])
```

```
x = np.array([1, 2, 3])
```

```
np.ones_like(x)
```

```
#array([1, 1, 1])
```

```

y = np.arange(10)
#[0 1 2 3 4 5 6 7 8 9]

np.linspace(-np.pi,np.pi,10)
#array([-3.14159265, -2.44346095, -1.74532925, -1.04719755, -0.34906585,
        0.34906585, 1.04719755, 1.74532925, 2.44346095, 3.14159265])

```

Operaciones básicas con arrays Las operaciones aritméticas pueden realizarse con ndarrays utilizando la misma sintaxis que la empleada para los escalares y evitando el uso de bucles explícitos.

```

data = np.array([[1, 2, 3], [4, 5, 6]])
data + 1
#array([[2, 3, 4], [5, 6, 7]])
data * 10
array([[10, 20, 30], [40, 50, 60]])

```

Arrays del mismo tamaño se pueden sumar y multiplicar celda a celda (si son de distinto tamaño no funciona):

```

data + data
#array([[ 2,  4,  6], [ 8, 10, 12]])
data * data
#array([[ 1,  4,  9], [16, 25, 36]])

```

Todas las operaciones booleanas (>, <, ==, etc) se aplican elemento a elemento y se devuelve un array de booleanos:

```

data = np.array([[1, 2, 3], [4, 5, 6], [4, 5, 6]])
#array([[1 2 3], [4 5 6], [4 5 6]])
data > 4
#array([[False, False, False], [False, True, True], [False, True, True]])

```

Ejercicios Construye una matriz de 5x5 elementos en la que los elementos de la diagonal sean 10 y los elementos fuera de la diagonal sean 5.

```

np.identity(5)*5+5
(np.identity(5)+1)*5

def diag2(n, f):
    ''' f: factor, n: dimensiones del array'''
    return (np.identity(n) + 1) * f
diag2(5,5)

```

Crea un array que contenga los días de la semana. ¿Cuál es su dtype?

```
weekdays_matrix = np.array(['Monday', 'Tuesday', 'Wednesday',
                             'Thursday', 'Friday', 'Saturday', 'Sunday'])
print(weekdays_matrix.dtype)
#<U9
#El 9 es la longitud de palabra más grande del array
```

1.2.1.2. Indexación y slicing

Para arrays unidimensionales, el funcionamiento es como en el caso de las listas:

```
x = np.arange(10)
#array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
x[4]
#4
x[2:7]
#array([2, 3, 4, 5, 6])
x[-3:]
#array([7, 8, 9])
```

Si se asigna un valor a un slice, el valor se asigna a todas las posiciones en el slice.

```
x[4:7] = -1
#array([ 0, 1, 2, 3, -1, -1, -1, 7, 8, 9])
x = np.arange(10)
#array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
x[0::2] = [19,17,15,13,11]
#array([19, 1, 17, 3, 15, 5, 13, 7, 11, 9])
```

Ten en cuenta que si obtienes una referencia a un slice, los elementos no se copian y sólo se mantiene la referencia a los elementos del slice. Esto significa que si después utilizas el slice para modificar sus valores, éstos se modifican en el array original. Si queremos una copia del slice y no una referencia a los valores en los arrays originales, se puede utilizar el método `copy()`.

```
x = np.ones((5,3), dtype=int)
#array([[1, 1, 1], [1, 1, 1], [1, 1, 1], [1, 1, 1], [1, 1, 1]])
y = x[:5:2,1:] #aquí se seleccionan primero filas y luego columnas
#array([[1, 1], [1, 1], [1, 1]])
y[:] = 25
print(x)
#array([[ 1, 25, 25], [ 1, 1, 1], [ 1, 25, 25], [ 1, 1, 1], [ 1, 25,
25]])

x = np.ones((5,3), dtype=int)
z = x[:3].copy()
#array([0, 1, 2])
y[2] = 25
print(x)
#[0 1 2 3 4 5 6 7 8 9]
print(y)
```

```
#[ 0 1 25]
```

En arrays multidimensionales, se utilizan comas para separar los slices para cada dimensión.

```
data = np.array([[1, 2, 3, 4], [5, 6, 7, 8], [9, 10, 11, 12], [13, 14, 15, 16]])
#array([[ 1,  2,  3,  4],
# [ 5,  6,  7,  8],
# [ 9, 10, 11, 12],
#[13, 14, 15, 16]])
data[1,1:3]
#array([6, 7])
data[1:3]
#array([[ 5,  6,  7,  8],
# [ 9, 10, 11, 12]])
data[:, 1:3]
#array([[ 2,  3],
# [ 6,  7],
# [10, 11],
# [14, 15]])
data[1:4:2, 1:4:2]
#array([[ 6,  8],
# [14, 16]])
```

Indexación booleana Es posible utilizar un array de valores booleanos (Verdadero/Falso) para seleccionar elementos del ndarray.

```
data = np.random.randn(10, 3) #generación de matriz 10x3 con números de
    distribución normal
data[data>0] #números positivos
ix = data[:,0] > 0 #índices booleanos para las filas en las que el
    primer elemento es positivo
#array([ True,  True, False, False,  True, False, False, False,
    True])
selected = data[ix, :] #seleccionar las filas que empiezan por un número
    positivo
```

La indexación booleana siempre crea una copia, a diferencia del slicing. Es posible combinar varias operaciones booleanas utilizando & (and), | (or) y ~ (negación). Además, se puede utilizar para asignar valores a los elementos de un array que cumplan con una condición.

```
data[(data[:, 0] > 0) & (data[:, 2] < 0), :] #selecciona las filas en
    las que el primer elemento es positivo y el último negativo
data[data < 0] = 0 #asigna el valor de 0 a todos los elementos negativos
```

Indexación fancy También es posible indexar utilizando matrices de enteros de forma similar al slicing. Estos enteros indican los índices de columna o fila a seleccionar. Este tipo de indexación es útil, por ejemplo, para seleccionar filas y columnas en un orden distinto al original.

```
nrows = 5
ncols = 3
p = 1
data = np.zeros((nrows, ncols))
for i in range(ncols):
    data[:, i] = np.arange(nrows)*p
    p*=10
#array([[ 0.,  0.,  0.],
#       [ 1., 10., 100.],
#       [ 2., 20., 200.],
#       [ 3., 30., 300.],
#       [ 4., 40., 400.]])
data[[3, 1]] #selección de las filas 3 y 1 en ese orden
#array([[ 3., 30., 300.],
#       [ 1., 10., 100.]])
data[:, [2, 0]]
#array([[ 0.,  0.],
#       [100.,  1.],
#       [200.,  2.],
#       [300.,  3.],
#       [400.,  4.]])
```

Ejercicios En el siguiente código, ¿qué diferencia supondría $y = 100 * y$ con $y[:]= 100*y$?

```
x = np.array([1, 2, 3, 4, 5])
y = x[:3]
y = 100*y
print(y) # [100 200 300]
print(x) # [1 2 3 4 5]
y = x[:3]
y[:] = 100*y
print(y) # [100 200 300]
print(x) # [100 200 300 4 5]
```

De ambas formas, el resultado para y es igual; la diferencia se encuentra en el array original x . En la primera variante, x no se ve modificado, mientras que en la segunda variante, los valores de x sí se han visto modificados.

El siguiente array contiene una lista de 100 notas aleatorias de 0 a 10. Crea un array que tenga una nota cualitativa para cada nota numérica:

```
n = 100
np.random.seed(13)
notas = np.random.rand(n)*10
```

```

notas_cual = np.array(notas, dtype=str)
notas_cual[notas < 5] = "SUSPENSO"
notas_cual[(notas >= 5) & (notas < 7)] = "APROBADO"
notas_cual[(notas >= 7) & (notas < 9)] = "NOTABLE"
notas_cual[notas >= 9] = "SOBRESALIENTE"

```

Ahora cuenta el número de suspensos, aprobados, notables y sobresalientes.

```

np.sum(notas_cual == "SUSPENSO")
np.sum(notas_cual == "APROBADO")
np.sum(notas_cual == "NOTABLE")
np.sum(notas_cual == "SOBRESALIENTE")

```

Hay más funciones de Numpy que no hemos visto. Por ejemplo, `np.unique` proporciona todos los valores únicos. Así, el ejercicio anterior se podría resolver de la siguiente forma:

```

for value in np.unique(notas_cual):
    print("# de ", value, ":", np.sum(notas_cual == value))

```

1.2.1.3. Operaciones con arrays

Transponer La función `transpose` o `T` permite intercambiar las filas por columnas de un array.

```

data = np.array([[1, 2, 3, 4], [5, 6, 7, 8], [9, 10, 11, 12], [13, 14, 15, 16]])
data.T # o data.transpose()

```

También es posible transponer arrays multidimensionales especificando los índices de los ejes que se quieren intercambiar.

Cambiar la forma de un array (reshape y ravel) El método `ndarray.reshape()` cambia la forma de un array sin cambiar su contenido, solo lo reajusta. El número de celdas totales en ambos arrays debería ser igual. El método `ndarray.ravel()` comprime un array en uno de una sola dimensión. Además, el método `np.concatenate` concatena dos arrays a lo largo de un eje, dado que el tamaño de la otra dimensión sea igual. Algunos ejemplos son:

```

x = np.arange(1, 13)
print(x)
y = x.reshape(3, 4) #cambia a dimensiones 3x4
print(y)
z = y.ravel() #vuelve a 1 dimensión
print(z)

x = np.zeros((3, 4))

```

```
print(x)
y = np.ones((3, 2))
print(y)
np.concatenate((x, y), axis = 1) #el eje 1 son las columnas
np.concatenate((x[:, :2], y), axis = 0) #eje 0 son las filas, y sólo se
    obtienen las primeras dos columnas de x
```

1.2.2. Matplotlib

Matplotlib es una librería de gráficos. Se suele importar como `import matplotlib.pyplot as plt`. La función más sencilla es `plot`.

```
x = np.linspace(0.1, 4, 40) #[0.1, 0.2, 0.3, 0.4, 0.5, ...]
y = 1/x #[10., 5., 3.3333, 2.5, 2., ...]
plt.plot(x, y)
```

También se pueden imprimir varias líneas a la vez:

```
x = np.linspace(-np.pi, np.pi, 40)
y1 = np.sin(x)
y2 = np.cos(x)
# Use dummy variable _ to get the return value of plot to avoid printing
    its reference
_ = plt.plot(x, y1, x, y2)
```

Opciones de trazado Se puede especificar la forma de trazado al proporcionar una cadena que incluya un identificador.

```
x = np.arange(0.1,4,0.1)
y = 1/x
plt.plot(x, y, 'm*')
```

En el caso anterior, `m` representa el color magenta y el asterisco la forma de cada punto en el gráfico. Así, las líneas discontinuas se muestran como dos guiones, guion punto alterna ambos símbolos, la `k` representa el color negro, etc. Todas las opciones se pueden encontrar en [la guía](#).

También se pueden retocar otras propiedades:

- `linewidth`: el ancho de la línea en píxeles
- `markeredgecolor`: el color del borde del marcador
- `markerfacecolor`: el color interno del marcador
- `markerzie`: el tamaño del marcador

En general, estos parámetros aplican a todo el gráfico. Sin embargo, se puede establecer anchos y colores distintos para cada una de las líneas llamando al `plot` varias veces:

```
x = np.arange(0.1,20,0.5)
y = np.exp(0.1*x)
y1 = y * np.sin(x)
y2 = y * np.cos(x)

_ = plt.plot (x, y1, ':ms', linewidth=2, markeredgecolor='k',
             markerfacecolor='r', markersize=5)
_ = plt.plot (x, y2, '-co', linewidth=11, markeredgecolor='b',
             markerfacecolor='c', markersize=11)
```

Tipos de gráficos Se pueden crear distintos tipos de gráficos:

- Gráfico de tarta (pie plot): mediante `plt.pie(x, Explode)`, siendo Explode la distancia a la que se debe separar la fracción.
- Histograma: `hist(x, nbins)`. Por defecto, se crean 10 cajas o bins equidistantes. Si `x` es una matriz, entonces se crea un histograma por columna.
- Diagrama de barras (bar plot): la función `bar(x, y, width, bottom, style...)` crea un gráfico de barras verticales u horizontales (`barh`). El parámetro `width` indica la separación entre las barras, y `bottom` permite poner barras apiladas.
- Contour: Esta función `contour(Z, n)` traza isolíneas desde la matriz `Z`, donde `Z` se interpreta como alturas con respecto al plano `xy`. `Z` debe ser al menos una matriz de `2x2`. El número de niveles puede especificarse utilizando el parámetro «`n`»; de lo contrario, se calculan automáticamente a partir de las alturas máxima y mínima de `Z`.
- Pcolor: La función `pcolor(c)` dibuja una matriz coloreada con colores dados por `C`. El gráfico muestra una columna y una fila menos que las columnas y filas de la matriz. Esta función (si no se dan otros parámetros) obtiene el color para la celda `(i, j)` del elemento `(i, j)` de la matriz `C` usando el mapa de colores actual.
- Diagrama de dispersión (scatter plot): La función `scatter(x, y, s = 20, c=u'b', ...)` crea un diagrama de dispersión con coordenadas `x` e `y`. Los vectores `x` e `y` deben tener el mismo tamaño. El tamaño de los puntos puede especificarse con `s` y su color con `c`. Tanto el tamaño como el color pueden darse para todos los puntos juntos utilizando un único valor o para cada punto individual utilizando un vector. Para la primera configuración, basta con dar un único escalar para el tamaño y una cadena para el color. Para la segunda, se debe dar un vector del mismo tamaño de `x` e `y` para los tamaños. Para los colores, se puede especificar un vector que contenga índices f el mapa de colores, o un array de tamaño `len(x)` x 3 con colores en RGB.

Opciones globales para gráficos Todos los gráficos admiten una serie de funciones que permiten crear un gráfico "decente":

- `figure`: crea una figura del tamaño especificado. En esta opción también se puede especificar el color de fondo.
- `subplot(nrows, ncols)`: crea una matriz de gráficos de filas `nrows` y columnas `ncols`. El último parámetro indica sobre qué gráfico vamos a trabajar. También hay más parámetros como `sharex` o `sharey` que permite crear ejes compartidos entre todos los subgráficos. También se debe indicar mediante índice el subgráfico que se quiere modificar a continuación.
- `xlabel ylabel`: permite indicar el nombre de los ejes `x` e `y`.
- `title`: permite especificar el título del gráfico.
- `legend`: añade la leyenda al gráfico. Se debe llamar esta opción una vez se haya ejecutado la función de creación del gráfico.
- `colorbar`: incluye una barra de colores de referencia.
- `xlim ylim`: modifica los límites de los ejes que se muestren en las gráficas.
- `xticks yticks`: modifica las marcas de los ejes.

Un ejemplo de como aplicar todo esto sería el siguiente (el resultado se ve en la figura 1.6):

```
plt.figure(num=None, figsize=(12, 4), facecolor='c')

# Data
x = np.linspace(-np.pi, np.pi, 20)
Y1 = np.sin(x)
Y2 = np.cos(x)

# Graph using legend and labels in the axis
plt.subplot(1,2,1) # Create a graph with two subplots. We start working
    on the first
plt.xlabel('x')
plt.ylabel('y')
plt.plot(x, Y1, label='sen')
plt.plot(x, Y2, label='cos')
plt.legend(loc=2) # Called after finishing the plot

# Graph with x label, title, with the limits of the axis modified
# and also the labels for the x axis
plt.subplot(1,2,2) # Now we use the second subplot
plt.xlabel('x')
plt.title('Seno y coseno')
plt.xlim([-3,3])
plt.xticks(range(-3,4), ['a', 'b', 'c', 'd', 'e', 'f', 'g'])
plt.plot(x, Y1, label='sen')
_ = plt.plot(x, Y2, label='cos')
```

Otro ejemplo sería:

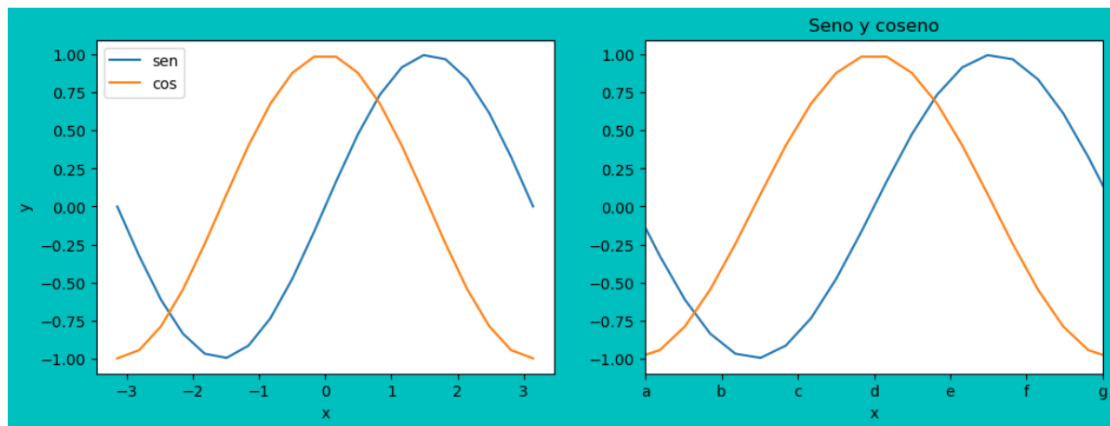


Figura I.6: Gráfico de ejemplo

```
def sigmoid(V, V0, W=12):
    """ Sigmoidal function

    Parameters
    -----
    V: float or array of floats
        Membrane potential
    V0: float
        Position of the sigmoid function
    W: float
        Width of the transition region

    Returns
    -----
    y: float or array of floats
        Sigmoidal function
    """
    return 1/(1+np.exp(-(V-V0)/W))

V = np.arange(-100, 100, 1) #if we use linspace, instead of the steps we
    put how many values we want
y = sigmoid(V, -40, 12)

fig, axs = plt.subplots(1, 2)
#varying V0
ax = axs[0]
for V0 in [-40, 0, 40]:
    y = sigmoid(V, V0, 12)
    ax.plot(V, y, label = f'{V0:.0f}mV')
ax.set_xlabel('V [mV]')
ax.legend(title = '$V_0$')

#Varying W
ax = axs[1]
for w in [1, 10, 20]:
```

```

y = sigmoid(V, 0, w)
ax.plot(V, y, label = f'{w:.0f}mV')
ax.set_xlabel('V [mV]')
ax.legend(title = '$w$')
plt.show()

```

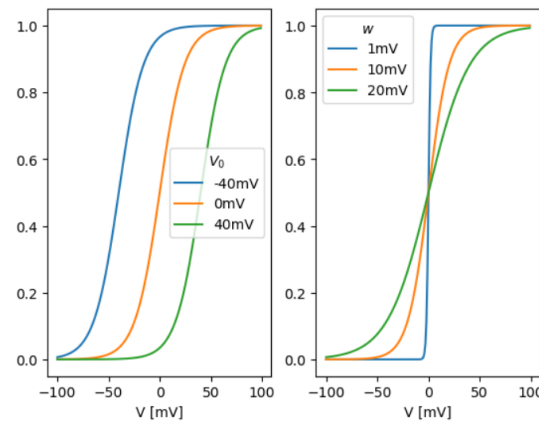


Figura I.7: Otro ejemplo

I.3. Biopython

Biopython es una librería Python con herramientas para computación biológica. Se trata de un proyecto de código abierto miembro de la Open Bioinformatics (OBF). Una fundación que promueve el código abierto en la investigación biológica. El objetivo de esta librería es facilitar al máximo el uso de python en bioinformática para el análisis sintáctico de archivos, funciones para trabajar con secuencias y características, trabajar con sitios bioinformáticos online como NCBI o ExPASy, alineamientos, etc.

I.3.1. Seq

La clase Seq es la clase básica para representar secuencias en Biopython. Se representan las secuencias biológicas como cadenas. Así, se puede iterar sobre las bases, realizar slicing, concatenar secuencias y utilizar otros métodos de cadenas. Todas estas operaciones devuelven nuevos objetos de tipo Seq, que son inmutables, como los strings.

Esta clase tiene métodos especiales de ADN como `complement` o `reverse_complement`.

```

dna = Seq('GATTACA')
print(dna.complement()) #CTAATGT
print(dna.reverse_complement()) #TGTAATC

```

También permite la transcripción mediante `transcribe` y `back_transcribe` para obtener el ARN y la traducción para obtener la secuencia de aminoácidos (`translate`). Se puede especificar la tabla de traducción y si se desea parar justo antes del primer codón de stop.

```

cds = Seq('GTGTTTTTGGTGTGGTGA')
mRNA = cds.transcribe()
print(mRNA) #GUGUUUUUGGUGUGGUGA
print(mRNA.back_transcribe()) #GTGTTTTTGGTGTGGTGA

mRNA = Seq('UUGUUUUUGGUGUGGUGA')

print("Translate using standard table: ", mRNA.translate()) #Translate
    using standard table: LFLVW*
print("Translate until stop codon : ", mRNA.translate(to_stop=True))
    #Translate until stop codon : LFLVW
print("Using a different table : ", mRNA.translate(table=2)) #Using a
    different table : LFLVWW
print("by name : ", mRNA.translate(table="Vertebrate
    Mitochondrial")) #by name : LFLVWW
# This next line raises an error because the sequence is not a CDS for
    table=2
#print("CDS-like RNA : ", mRNA.translate(table=2,
    cds=True)) #CDS-like RNA : MFLVW
# But it is a CDS-like sequence for table=11 (Bacterial)
print("CDS-like RNA : ", mRNA.translate(table=11,
    cds=True)) #CDS-like RNA : MFLVW

cds = Seq('GTGTTTTTGGTGTGGTGA')
prot = cds.translate(table=11, cds=True)
prot = cds.translate()
prot #Seq('VFLVW*')

```

Si la secuencia es un CDS, debe tener un codón de inicio y final; si no, se produce un error. No es necesario seguir el flujo natural: se pueden obtener los aminoácidos directamente desde el ADN sin tener que pasar previamente por el ARN.

I.3.1.1. Tablas de traducción

Se pueden obtener las tablas de traducción:

```

from Bio.Data import CodonTable

bacterial_table = CodonTable.unambiguous_dna_by_name['Bacterial']
print(bacterial_table)

```

I.3.2. SeqRecord

SeqRecord consta de una secuencia (objeto Seq) junto con id, descripción, anotaciones, características, etc. Se puede crear un SeqRecord a partir de un objeto Seq, de un archivo fasta, genbank, etc.

```

from Bio.SeqRecord import SeqRecord

```

```
sr = SeqRecord(Seq('AAA'), id='1', description='Simple seq',
               annotations={"molecule_type": "DNA"})
print(sr)
```

1.3.3. SeqIO

IO viene de Input Output. Bio.SeqIO proporciona acceso a la mayoría de los formatos de archivos bioinformáticos. La función `read()` carga un único archivo de secuencia. Obtiene como parámetros el nombre y formato del archivo. Si hay más de un registro esta función genera un error. Esta función devuelve un `SeqRecord`.

```
from Bio import SeqIO

record = SeqIO.read("phix174/phix.fa", "fasta")
print(record)
```

La función `parse` recupera varias secuencias de un archivo. Los parámetros son los mismos que `read()`: Nombre y formato del fichero. Esta función devuelve un iterador que permite recorrer todas las secuencias y puede iterar con un bucle

```
for record in SeqIO.parse("other/ls_orchid.fasta", "fasta"):
    print(record)
    print()
```

Otra función es `to_dict`, la cual crea un diccionario del iterador con todas las secuencias del fichero. El problema es que se guarda en la memoria, por lo que puede ralentizar el ordenador en caso de ficheros grandes.

```
iterator = SeqIO.parse("other/ls_orchid.gb", "genbank")
records_dict = SeqIO.to_dict(iterator)
print(records_dict['Z78533.1'])
```

La función `index()` crea un pseudo diccionario a partir de un fichero. Admite como parámetros el nombre y el formato del archivo. Esta función no almacena todo en memoria, sino que obtiene la información del disco cuando es necesario. Almacena la posición de cada secuencia en el fichero para un acceso rápido. Desde el punto de vista del usuario es como un diccionario de secuencias. Es importante cerrar el fichero cuando ya no se utilice.

```
records_dict = SeqIO.index("other/ls_orchid.gb", "genbank")
print(records_dict['Z78533.1'])
records_dict.close()
```

También se puede trabajar con ficheros comprimidos para ahorrar espacio en el disco:

```
import gzip
```

```
with gzip.open("arabidopsis_thaliana/GCF_000001735.3_TAIR10_rna.fna.gz",
    "rt") as f:
    total_len = 0
    for sr in SeqIO.parse(f, "fasta"):
        total_len += len(sr.seq)
    print(total_len)
```

Trabajar con índices y archivos comprimidos es más complicado, ya que la posición de cada secuencia no es fácil de obtener. El formato Blocked GNU Zip Format ofrece una buena solución. Comprime los datos por bloques para permitir la indexación. La función `Index()` puede trabajar con el formato de archivo BGZF.

I.3.4. Atributos de SeqRecord

Los atributos más importantes de SeqRecord son:

- `id` y `nombre`: identificador y nombre de la secuencia
- `seq`: objeto Seq que contiene la secuencia actual
- `annotations`: diccionario de Python con toda la información de la secuencia. Se pueden encontrar tipos de moléculas, topología, fecha, accessiones, versión de secuencia, `gi`, palabras clave, fuente, organismo, taxonomía, referencias, etc.
- `features`: lista de un objeto SeqFeature que contiene la información sobre las distintas características encontrada en una secuencia.

```
records_dict = SeqIO.index("other/ls_orchid.gbk", "genbank")
record = records_dict['Z78533.1']
records_dict.close()

print("ID:", record.id, "Name:", record.name)
print("Description:", record.description)
print(record.annotations)
print(record.features)
```

I.3.4.1. SeqFeature: características de SeqRecord

SeqFeature es una clase que incluye información sobre las características de una secuencia. Sus atributos principales son:

- `type`: descripción de la característica. Puede ser "gene", "CDS", "region", etc. Es similar al campo 3 de un fichero gff.
- `location`: región donde se encuentra la característica. Contiene las posiciones de inicio y final. Este atributo también puede manejar características localizadas en múltiples regiones.

- **qualifiers**: un diccionario de Python con información sobre la característica.

```
from Bio import SeqFeature

records_dict = SeqIO.index("other/ls_orchid.gbk", "genbank")
print(len(records_dict['Z78533.1'].features), "features found")

for feature in records_dict['Z78533.1'].features:
    print("-----")
    print(feature)

records_dict.close()
```

Esta información está incluida en SeqRecord cuando se lee un fichero que contenga la información, como es el caso de GenBank. Por ejemplo, si se abre un fichero FASTA, SeqRecord no encontrará ninguna característica.

Operador in El operador `in` se utiliza para comprobar si una posición concreta se encuentra en una característica. Por ejemplo, el siguiente código imprime la característica que incluye la posición 10000.

```
record =
    SeqIO.read('haemophilus_influenzae/GCF_000027305.1_ASM2730v1_genomic.gbff',
              'genbank')

position = 10000
for feature in record.features:
    if position in feature:
        print(feature)
        print('-----')
```

Este operador también funciona con objetos Location.

1.3.4.2. SeqFeature: método extract

El método `extract()` extrae las características de la secuencia completa. El siguiente código itera sobre las secuencias de un fichero y después sobre las características de cada secuencia, extrayendo las características CDS y generando la proteína asociada utilizando la tabla de traducción correspondiente.

```
n_cds = 0
file_iter =
    SeqIO.parse('haemophilus_influenzae/GCF_000027305.1_ASM2730v1_genomic.gbff',
              'genbank')
#file_iter =
    SeqIO.parse('plamodium_falciparum/GCF_000002765.3_ASM276v1_genomic.gbff',
              'genbank')

for seqrec in file_iter:
```

```

for feature in seqrec.features:
    if feature.type == 'CDS':
        n_cds += 1
        mRNA = feature.extract(seqrec)
        try:
            transl_table = 1
            if 'transl_table' in feature.qualifiers:
                transl_table =
                    int(feature.qualifiers['transl_table'][0])
            p = mRNA.seq.translate(table=transl_table, cds=True)
        except:
            print("Protein {0} in gene {1} could not be translated!".
                format(feature.qualifiers['protein_id'][0],
                    seqrec.id))

print(n_cds)

```

I.3.4.3. SeqRecord: slicing

Es posible utilizar slicing con SeqRecord. El SeqRecord que se devuelve contiene solo las características que se incluyen en la subsecuencia. Las anotaciones no se copian en las subsecuencias.

```

record =
    SeqIO.read('haemophilus_influenzae/GCF_000027305.1_ASM2730v1_genomic.gbff',
        'genbank')
print(record.id)
print("# of feature: ", len(record.features))
print("# of annotations: ", len(record.annotations))
print(record.features[1200])

sub_record = record[612623:620000]
print(sub_record.id)
print("# of feature: ", len(sub_record.features))
print("# of annotations: ", len(sub_record.annotations))

print(sub_record.features[0])

```

Capítulo II

Acceso programático a bases de datos biomédicas

II.1. Introducción

En este bloque queremos acceder mediante los programas a las bases de datos biomédicas y poder interactuar con ellas. Para las bases de datos biomédicas hay que hacer peticiones desde el protocolo HTTP. Un protocolo es una forma de interacción entre dos partes: queremos hacer una petición a la base de datos a partir de una entrada para que nos devuelva una respuesta que podamos utilizar posteriormente. Tenemos que usar los protocolos intermediarios para poder interactuar con la base de datos. Como esto puede ser un lío, se diseñaron las API REST. Un API es application programming interface, es decir, un protocolo o una manera de interactuar con otra parte en un formato concreto. REST tiene que ver con el protocolo HTTP de base. El protocolo HTTP se utiliza para navegar por internet. Muchas aplicaciones exponen un API REST, lo que quiere decir que hay unas ciertas órdenes o funciones que se pueden llamar para interactuar con la aplicación. Hay una serie de endpoints, que son cada una de las funciones que están disponibles para llamar: un endpoint para buscar un organismo concreto, para descargar una cierta secuencia, etc.

Hay varias maneras de acceder a bases de datos en línea (nivel de abstracción o automatización ascendente):

- **Copiar la base de datos de manera local:** esto es un proceso manual que hace que la base de datos se quede rápidamente desactualizada. Se puede descargar un fichero de un servicio web en Linux mediante `wget url/base/de/datos/fichero.txt`
- **Rellenar formularios:** todavía el sistema no ha dado acceso al público a sus servicios mediante API, pero se puede escribir en Python un programa que rellene el formulario como si fuera un humano para acceder a los datos cambiando los valores de los parámetros. En general, la sintaxis de una URL es la siguiente:

`scheme:[//host]path[?query][#fragment]`

siendo `scheme` HTTP o HTTPS cuando el protocolo está cifrado y es seguro (la `s` viene de `secure`), `host` el servidor que contiene la base de datos a la

que se quiere acceder, el path la ruta para llegar a una parte del servidor (como si fueran carpetas) y la interrogación el marcador de los argumentos (las queries se escriben en formato nombre=valor&nombre2=valor2). Así, lo que queremos hacer con Python es codificar los distintos argumentos para poder acceder a varias entradas de forma automática. Por ejemplo, en el formulario tipo https://www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=ena_sequence&id=J00231&style=raw, en lugar de rellenar el formulario varias veces, queremos cambiar el valor de los distintos argumentos (id=J00231, id=J00232, id=J00233, db=afdb, etc.).

- **Peticiones HTTP directas:** esto se puede conseguir mediante la librería requests en Python o mediante la terminal. Se emplean las URLs de los distintos servidores con métodos como GET y POST.
- **Uso específico de API REST:** Hay algunos servicios que proporcionan APIs de los servidores para poder utilizarlos desde Python. Esto se conoce como APIs de alto nivel o SDKs. Esta es la mejor manera, ya que es más cómoda y más fácil de utilizar, pero no está disponible en todos los sistemas. Estas API REST están muy encapsuladas, por lo que no disponen de URLs ni GET o POST.

II.1.1. Acceso programático en formularios

Hay varias librerías disponibles en Python para automatizar el acceso y procesamiento de las URL:

- `urllib`: acceso de bajo nivel, más adecuado para programación de red.
- `requests`

La librería requests tiene el comando `get`, que es el método HTTP más común.

```
import requests
url =
    "https://www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=ena_sequence&id=J00231&style=raw"
response = requests.get(url)
print(response.text)
```

En el método `get` de requests se puede utilizar el argumento `params` para pasar los distintos argumentos. Este argumento debe ser un diccionario con los distintos parámetros:

```
import requests

ebi_url = 'https://www.ebi.ac.uk/Tools/dbfetch/dbfetch'

response = requests.get(ebi_url,
    params={'db': 'ena_sequence', 'id': 'J00231', 'style': 'raw'})

print(response.text)
```

II.2. JSON

Hay distintas formas de poder representar un fichero:

- Texto plano
- HTML (HyperText Markup Language)
- XML (eXtensible Markup Language)
- JSON (JavaScript Object Notation)

La notación JSON es similar a la de los diccionarios en Python conceptualmente: las claves son strings y los valores pueden ser strings, números, listas o incluso diccionarios. Técnicamente es una forma de estructurar la información (figura II.1) en formato ASCII.



Figura II.1: Ejemplo de un fichero JSON que describe a una persona. Las llaves representan estructuras y los corchetes listas.

Python cuenta con una librería llamada json.

```
import json

json_data = '{"name": "John", "age": 30, "city": "New York"}'

print(json_data)
```

Para cargar un JSON, se utiliza el método `json.loads()`. Así, el JSON se guarda en un diccionario que posteriormente se puede guardar serializado en disco. Otra función es `json.dumps()`, el cual coge un diccionario en una cadena tipo JSON. La función `json.dump` escribe un objeto JSON en un fichero.

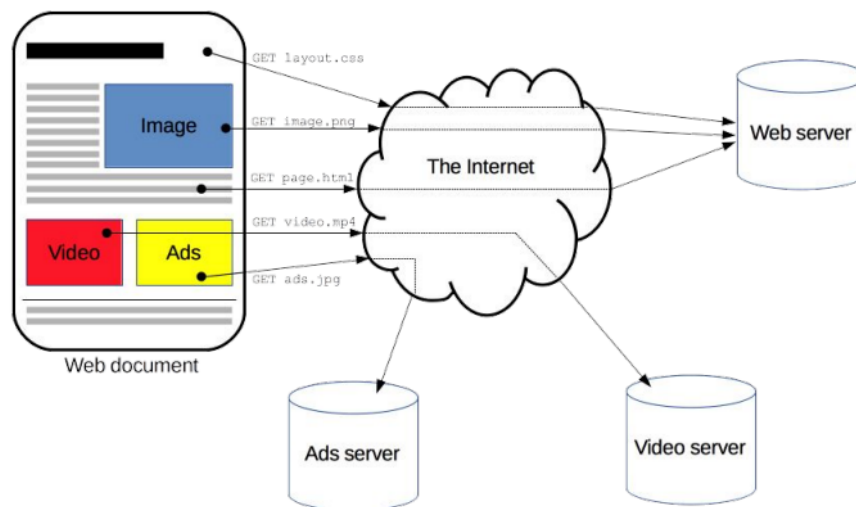
<code>dict = json.loads(str)</code>	Convierte un string en un diccionario
<code>str = json.dumps(dict)</code>	Convierte un diccionario en un string
<code>json.dump(dict, file)</code>	Escribe un objeto JSON (dict) en un fichero)

Tabla II.1: Resumen

II.3. Protocolo HTTP

Los protocolos de acceso a las bases de datos se han montado sobre el protocolo HTTP. HTTP es un protocolo para obtener recursos en la web, como documentos HTML, imágenes y otros. Es un **protocolo cliente-servidor**, lo que significa que las peticiones las inicia el destinatario, normalmente el navegador web.

Una página web es una colección de recursos: imágenes, vídeos, anuncios, etc. Estos elementos pueden estar en sitios distintos, y cuando la página se carga, se embebe en el código y se muestra.



Los clientes y los servidores se comunican intercambiando mensajes individuales (a diferencia de un flujo de datos). Los mensajes enviados por el cliente, normalmente un navegador web, se denominan **peticiones (requests)** y los mensajes enviados por el servidor como respuesta se denominan **respuestas (responses)**.

II.3.1. Aspectos básicos de HTTP

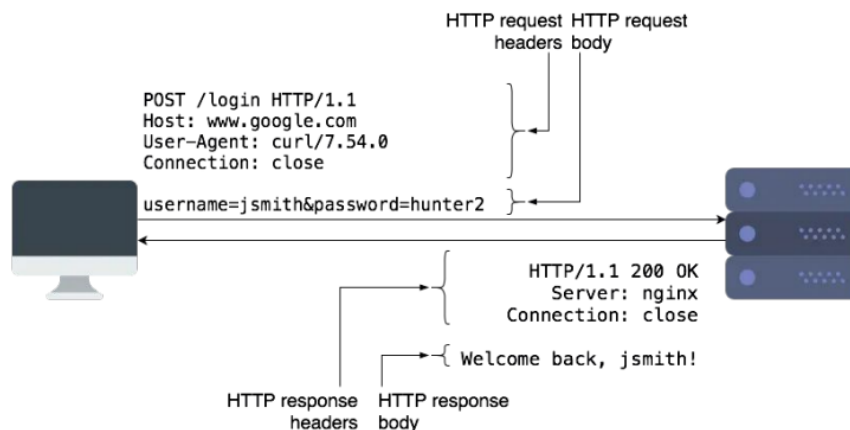
HTTP se diseñó para ser sencillo y legible para los humanos. Los mensajes HTTP pueden ser leídos y comprendidos por humanos, lo que facilita las pruebas a los desarrolladores y reduce la complejidad para los recién llegados. Además, HTTP es extensible: Las cabeceras HTTP hacen que este protocolo sea fácil de ampliar y experimentar. Se pueden introducir nuevas funciones mediante un simple acuerdo entre un cliente y un servidor sobre la semántica de una nueva cabecera.

HTTP no tiene estado, es decir, no existe ningún vínculo entre dos solicitudes que se realizan sucesivamente en la misma conexión. Las cookies HTTP permiten el uso de sesiones con estado (por ejemplo, para utilizar cestas de la compra de comercio electrónico).

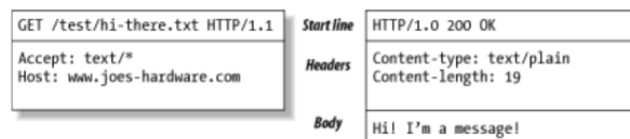
II.3.2. Flujo HTTP

Cuando un cliente quiere comunicarse con un servidor realiza los siguientes pasos:

1. Abre una conexión TCP: la conexión se utiliza para enviar una petición, o varias, y recibir una respuesta.
2. Envía un mensaje HTTP
3. Espera una respuesta



Las peticiones son puro texto (no hay datos binarios). La estructura es la siguiente: verbo, url y cabecera para proporcionar información adicional (por ejemplo, el idioma de una página web). La respuesta contiene una línea inicial con el HTTP y un código asociado (200 si todo está bien, 404 si no se ha encontrado, etc), la cabecera y el cuerpo.



Los códigos de retorno son los siguientes:

Rango disponible	Rango definido	Categoría
100-199	100-101	Información
200-299	200-206	Éxito
300-399	300-305	Redirección
400-499	400-415	Error del cliente
500-599	500-505	Error del servidor

Tabla II.2: *Códigos de retorno*

Un servidor es un ordenador en un data center. Cada servicio tiene asignado un puerto, ya que en un mismo servidor puede haber varios servicios web. Las direcciones IP (técnicamente IPv4) se dividen en cuatro bloques de tres números que van del 0 al 255 (esto ocupa 8 bites). Así, en el mundo puede haber un máximo de 255^4 direcciones IP en el mundo. Por ello, a los servicios se les añade después de la dirección IP el puerto. En el caso de la web de la UAM, se trata de 150.244.214.237:80 para

HTTP y 150.244.214.237:443 para HTTP. Todos los servicios seguros (HTTPS) se conectan al puerto 443. Una vez conectados al servidor, se puede utilizar el método GET para obtener como respuesta la web que se está pidiendo. Esto se puede realizar en Python con la librería requests y la función get, como se explicó anteriormente. También está el método POST para enviar datos o información a un servidor. Estos datos se incluyen en el cuerpo de la petición, no están visibles en la URL, por lo que se suele utilizar para los login al ser más seguro (usuario y contraseña está oculto a la vista). En resumen: se pide información con GET y se envía información con POST.

Ejemplo: <https://rest.ensembl.org/> Aquí se pueden ver todos los endpoints (las distintas funcionalidades) de la página Ensembl. Con GET archive/id/:id se puede acceder a la última versión del identificador que se dé. id indica la variable que se debería proporcionar, es decir, sustituirlo con el ID. Lo que se devuelve está en formato JSON, pero al leerlo en Python será del tipo string. En este ejemplo, hay un endpoint POST archive/id que lo que hace es devolver la última versión para un conjunto de identificadores. Antes hemos dicho que POST se utiliza para enviar información y GET para obtenerla, pero esta API está mal diseñada y diferencia GET y POST en que el primero es para un solo identificador y el segundo para varios.

En esta página se proponen ejemplos de peticiones. Para GET sería la siguiente:

```
import requests, sys

server = "https://rest.ensembl.org"
ext = "/archive/id/ENSG00000157764?"

r = requests.get(server+ext, headers={ "Content-type" :
    "application/json"})

if not r.ok:
    r.raise_for_status()
    sys.exit()

decoded = r.json() #sería mejor utilizar json.loads
print(repr(decoded))
```

II.3.3. APIs con HTTP

Los métodos HTTP originales tienen la siguiente semántica:

- GET: pide obtener una representación del recurso identificado
- POST: pide añadir información al recurso
- HEAD: como GET, pero sin recibir el cuerpo, solo la cabecera
- Otros verbos como DELETE, PUT, CONNECT, TRACE, etc.

Para nuestros fines, se utilizará casi siempre GET, pero algunos servidores de datos biomédicos también aceptan POST. Sin embargo, algunos servidores que aceptan

POST no lo utilizan para el fin previsto (crear información en el servidor), sino sólo para responder con la representación del recurso (al igual que GET). URLs como las siguientes utilizan el método GET:

<https://rest.ensembl.org/sequence/id/ENSG00000157764>

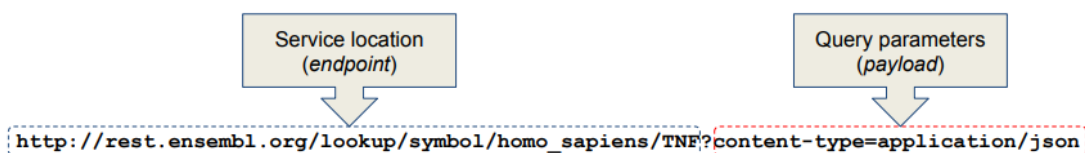
https://www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=uniprot&id=WAP_RAT

Un método POST equivalente utiliza una URL sin parámetros (<http://rest.ensembl.org/sequence/id/>) que se deben enviar junto a los parámetros en el cuerpo de la petición (`"ids":["ENSG00000157764"]`).

```
## GET ##
import requests
gene = "ENSG00000157764"
endpoint = f"http://rest.ensembl.org/sequence/id/{gene}"
r = requests.get(endpoint, headers={ "Content-Type" : "text/plain"})

#Especificando params
gene = "ENSG00000157764"
endpoint = f"http://rest.ensembl.org/sequence/id/{gene}"
parameters = {'species': 'homo_sapiens'}
r = requests.get(endpoint, params = parameters, headers={"Content-type"
: "text/json"})

## POST ##
#Especificando data en lugar de params
response = requests.post(server, data = {'key': 'value'})
```



Aunque las bases de datos sean de acceso público (open access), usualmente piden una API Key, que es un número largo único para cada usuario registrado que se pone en la cabecera. De esa forma, proporcionando el API Key, el servidor sabe qué usuario se está conectando y puede limitar las peticiones de un mismo usuario.

II.4. API REST

REST viene de REpresentational State Transfer. Realmente no es un protocolo, si no un estilo de prestación de servicios web basados en HTTP. Los **servidores** aportan representaciones textuales de recursos (como XML, JSON o HTML), y los **clientes** piden servicios mediante los métodos HTTP (GET, POST, etc). El estado no se guarda en el servidor (cada petición se responde de forma independiente sin memoria de peticiones anteriores).

II.5. Examen de prueba

II.5.1. Ejercicio 1

(1,5 puntos) Implementa una función `write_dna_sequece` que recupere de `rest.ensembl.org` y escriba en un fichero, cuyo nombre se reciba como parámetro, la secuencia ADN de un identificador dado. Adicionalmente, si se especifica un parámetro opcional a la función denominado `'analysis'` y éste tiene el valor `True`, la función deben imprimir también por pantalla el número total de bases, así como el número de cada una de ellas. La salida debe ser similar a la siguiente:

File 'dna_sequence.txt' successfully written.

Total number of bases: 205603

Number of 'A' bases: 60015

Number of 'C' bases: 37965

Number of 'G' bases: 40168

Number of 'T' bases: 67455

Para la implementación de la función, puedes utilizar el siguiente endpoint:

GET `/sequence/id/:id` recupera secuencias por identificador.

cuya documentación se muestra más abajo. No es necesario el uso de parámetros opcionales del endpoint.

GET sequence/id/:id

Request multiple types of sequence by stable identifier. Supports feature masking and expand options.

Parameters

Required

Name	Type	Description	Default	Example Values
id	String	An Ensembl stable ID	-	ENSG00000157764 ENSG00000157764.fasta (supported on some deployments)

Optional

Resource Information

Methods	GET
Response formats	fasta json seqxml text yaml jsonp
Slice length	1e7

Figura II.2: Documentación del endpoint

```
import requests

def write_dna_sequence(DNA_id, file_name, analysis = False):
    """Función que obtiene la secuencia de Ensembl y la escribe en un
        fichero.
        Input Parameters
        -----
        DNA_id: str
            ID de la secuencia que se quiere guardar
        file_name: str
            nombre del fichero en el que se quiere guardar la secuencia
        analysis: bool
            Parámetro que imprime la cantidad total de bases y el número de
            bases de cada nucleótido cuando True (por defecto).
    """
    server = "rest.ensembl.org/"
    endpoint = f"sequence/id/{DNA_id}"
```

```

response = requests.get(f"{server}{endpoint}", headers =
    {"Content-type": "application/text"})

if response.ok:
    dna_sequence = response.text
    with open(file_out, "w") as f:
        f.write(dna_sequence)
    print(f"File {file_out} successfully written.")

    if analysis:
        print(f"Total number of bases: {len(dna_sequence)}")
        for nucleotide in "ATCG":
            print(f"Number of {nucleotide} bases:
                {dna_sequence.count(nucleotide)}")
    else:
        print("Se ha producido un error")

```

II.5.2. Ejercicio 2

(1 punto) El siguiente programa tiene como objetivo obtener la información de la versión más reciente de un gen específico. Desafortunadamente, no funciona como se espera. Señala y explica todos los errores, y escribe una versión corregida.

```

import requests, sys, json
import pprint

server = "http://rest.ensembl.org"
gene = "ENSG00000157764"
get_gene_info_endpoint = f"/overlap/id/{gene}?feature=gene"
r = requests.get(f"{server}{get_gene_info_endpoint}")

if r.ok:
    r.raise_for_status()
    sys.exit()

pprint(r)

```

Primero, la comprobación de la petición está invertida: cuando la petición se ha realizado correctamente, queremos continuar con el programa, y cuando ha habido algún error, se quiere apagar el sistema. Además, como importamos el módulo pprint, al llamar a una función dentro de ese módulo se debe utilizar la estructura módulo.función, no solo la función. Asimismo, como queremos imprimir el contenido de la petición, esto se encuentra dentro de text. Finalmente, se está importando el módulo json y no se está utilizando. Esto no es un error en sí, solo un warning, pero se puede reemplazar pprint con json.loads(r.text) y un print normal. Así, el código corregido quedaría de la siguiente forma:

```

import requests, sys, json

```

```
import pprint

server = "http://rest.ensembl.org"
gene = "ENSG00000157764"
get_gene_info_endpoint = f"/overlap/id/{gene}?feature=gene"
r = requests.get(f"{server}{get_gene_info_endpoint}")

if not r.ok: #r.ok == False
    r.raise_for_status()
    sys.exit()

pprint.pprint(r.text)

##Alternativa
r_dict = json.loads(r.text)
print(r_dict)
```

Capítulo III

La línea de comandos de Linux

III.1. Introducción

III.1.1. Test inicial

La línea de comandos de Linux se puede abrir mediante la combinación de teclas Ctrl + Alt + t. Ejecutamos el siguiente comando: `wget archive.ics.uci.edu/ml/machine-learning-datasets/adult/adult.data`. El comando `wget` sirve para la descarga no interactiva de ficheros desde la web, soportando los protocolos HTTP, HTTPS y FTP. Con dicho comando, nos descargamos desde la web el fichero `adult.data` que tiene un tamaño de 3,8 M con permisos de lectura y escritura para el usuario y grupo y solo de lectura para otros. Este fichero contiene datos demográficos de distintas personas, incluyendo la edad, empleo, educación, estado civil, etnia, sexo y país.

```
#Valor del cuarto campo de la tercera fila
```

```
cut -d "," -f 4 adult.data | head -n 3 | tail -n 1 #HS-grad
```

```
#Cuántas líneas contienen Portugal
```

```
grep -c "Portugal" adult.data #37
```

```
#Línea en la que aparece Portugal por primera vez
```

```
grep -n "Portugal" adult.data | head -n 1 #360
```

```
#Valor del primer campo en la fila 1000 empezando por el final
```

```
cut -d "," -f 1 adult.data | tail -n 1000 | head -n 1 #40
```

```
#Contar el número de filas del fichero
```

```
wc -l adult.data #32562
```

```
#Crear un fichero con las primeras 1000 filas
```

```
head -n 1000 adult.data > adult-first-1000.data
```

```
#Contar bytes del fichero nuevo
```

```
wc -c adult-first-1000.data #121895
```

```
#Contar las líneas que contienen el string "Married-civ-spouse"
```

```
grep -c "Married-civ-spouse" adult.data #14976
```

```
#Valor mayor y menor del primer campo
```

```
cut -d "," -f 1 adult.data | sort -n | tail -n 1 #90
```

```
cut -d "," -f 1 adult.data | sort -n | head -n 2 #17, hay un blanco
```

III.1.2. Redirección y tuberías

Se puede redireccionar la salida a un fichero mediante `>` y `»` para sobrescribirlo o añadirlo respectivamente.

Con algunas acciones, la línea de comandos se queda bloqueada hasta que termine la tarea (por ejemplo, con `gedit`). Con `Ctrl C` se cierra el procedimiento que estaba corriendo y se recupera el control de la consola de comandos, pero los cambios sin guardar se pierden. Por tanto, se puede enviar ese trabajo al background mediante `&`. Otra opción es `Ctrl Z`, que pausa el procedimiento y permite recuperar el control de la consola.

Las tuberías permiten redireccionar la salida de un comando como input de otro sin necesidad de crear ficheros. Estas tuberías pueden ser concatenadas.

III.1.3. Filtros

Un filtro es un programa que lee la entrada estándar, realiza alguna operación sobre ella y saca el resultado por la salida estándar. Normalmente, se combinan con tuberías, y algunos comandos que pueden servir como filtros son `head`, `tail`, `tr`, `fmt`, `grep`, `sort`.

tr El comando `tr` sirve para traducir o reemplazar. Por ejemplo, `tr '[0-9]' '*'` reemplaza todos los números por asteriscos. Si se utiliza `-d`, se borra el parámetro que se le dé en lugar de reemplazarlo. Asimismo, `-c` sirve para reemplazar todo lo que no entre en el filtro a lo que se le pase como segundo parámetro. Otro argumento opcional es `-s`, que comprime una serie de caracteres repetidos.

head Los comandos `head` y `tail` permiten mostrar las primeras y últimas líneas respectivamente, pero también tienen algunas opciones adicionales. Cuando se utiliza `-n -100`, se excluyen las últimas 100 líneas, mientras que `-n +100` hace que empiece en la fila 100.

fmt El comando `fmt` permite formatear ficheros para que las líneas sean más pequeñas, recolocar palabras, etc. Por ejemplo, `fmt -70 -s` hace que cada línea muestre como máximo 70 caracteres, mostrando los caracteres restantes en una nueva línea.

```
#Muestra los 10 procesos más recientes del sistema del usuario actual
```

```
ps -ef | tr -s ' ' | grep 'sandra ' | tail -n 10
```

```
#Muestra las páginas del manual de grep con 100 caracteres por línea,  
sustituyendo los dígitos por asteriscos y redireccionando a un fichero  
man grep | fmt -100 | tr '[0-9]' '*' > file.txt
```

split El comando `split` divide un fichero en distintos bloques de 1000 líneas de forma predeterminada. Ese valor se puede modificar con `-l`, o se puede definir la cantidad de ficheros de tamaño equitativo sin romper líneas con `-n`.

cut El comando `cut` borra secciones de cada línea de un fichero. Con `-d` se define el delimitador de las distintas columnas, y con `-f` se selecciona la columna.

paste El comando `paste` permite combinar varios ficheros que consistan de distintas columnas. Se puede definir el delimitador con `-d`, que de forma determinada utiliza tabulador. Así, es el comando contrario a `cut`.

#Extrae las columnas 6 y 4 del fichero `adult.data` en ese orden.

```
cat adult.data | cut -d ',' -f 6 > adult-f6.data
```

```
cat adult.data | cut -d ',' -f 4 > adult-f4.data
```

```
paste -d ',' adult-f6.data adult-f4.data > adult-f6-f4.data
```

Esto puede ser útil para combinar un fichero sin numerar con una columna que contenga la numeración.

```
wc -l adult.data #32562
```

```
echo {1..32562} | tr ' ' '\n' > column-numbers.txt
```

```
paste -d ',' column-numbers.txt adult.data > adult-numbered.data
```

shuf El comando `shuf` (de shuffle) permite generar permutaciones aleatorias, mezclando o barajando las líneas. Con el parámetro `-r`, se permiten las repeticiones, y `-n` permite especificar el número de mezclas.

#Mezclar las primeras 10 filas

```
head adult-numbered.data | shuf
```

#Mezclar las primeras 5 filas

```
head adult-numbered.data | shuf -n 5
```

#Generar 15 filas mezcladas con repetición

```
head adult-numbered.data | shuf -n 15 -r
```

#Crear una permutación aleatoria de los números 2 a 9

```
shuf -i 2-9
```

sort El comando `sort` ordena las filas utilizando ordenación alfabética ASCII (las mayúsculas están por delante de las minúsculas; Z va antes que a). Para ignorar las mayúsculas y minúsculas, se debe poner `-f`, y para ordenación numérica, `-n` o `-g`.

uniq El comando `uniq` elimina las filas adyacentes repetidas. Esto se suele utilizar junto a `sort` para primero ordenar las filas y luego borrar las repetidas. `-c` indica cuántas filas había de cada tipo.

#Extraer el tipo de región de un fichero GFF y contar cuántas filas hay de cada tipo

```
cut -f 3 GCF_000002765.3_ASM276v1_genomic.gff | grep -v '#' | sort |  
uniq -c
```

#Extraer una lista ordenada de todos los posibles valores del campo 6 de `adult.data`

```
cut -f 6 -d , adult.data | sort | uniq -c
```

join

El comando join une líneas de dos ficheros en base a un campo en común. -1 sirve para especificar el campo del fichero 1, y -2 igual para el segundo fichero. -t permite especificar el delimitador.

```
cut adult-numbered.data -d ',' -f 1,3 | shuf -n 1000 | sort -t ',' -k
1,1 > f13.txt
cut adult-numbered.data -d ',' -f 1,7 | shuf -n 1000 | sort -t ',' -k
1,1 > f17.txt
join -t ',' -1 1 -2 1 f13.txt f17.txt
#Y lo mismo pero manteniendo todas las líneas del primer fichero
join -t ',' -1 1 -2 1 -a 1 f13.txt f17.txt
```

comm

El comando comm sirve para comparar dos ficheros ordenados de forma alfabética. El resultado son tres columnas. En la primera, se muestran los datos únicos del primer fichero, en la segunda, los datos únicos del segundo, y en la tercera, los datos comunes en ambos ficheros.

```
#Crea un fichero a.txt con 20 números aleatorios entre 0 y 100 sin
repetir; cada número debe estar en una línea distinta
shuf -i 0-100 -n 20 | sort > a.txt
#Crea un segundo fichero b.txt con 20 números aleatorios entre 0 y 100
sin repetir; cada número debe estar en una línea distinta
shuf -i 0-100 -n 20 | sort > b.txt
#Muestra los números que aparezcan en ambos ficheros
join -t ',' -1 1 -2 2 a.txt b.txt
comm -12 a.txt b.txt
```

III.1.4. Scripts de shell

Todos los comandos vistos anteriormente se pueden utilizar en scripts de bash. Los script deben empezar con el intérprete.

```
#!/bin/bash
#Hello world script

echo "Hello World!"
```

Para ejecutar un script, primero se le debe dar permisos de ejecución mediante `chmod +x hello-world.sh`. Dentro de un fichero se pueden crear variables asignando el valor con el igual sin dejar espacios. Para luego poder llamar a la variable, se debe llamar con un `$` delante del nombre.

```
#!/bin/bash
#Variables:
input_file='adult.data'
output_file='adult5-7.data'
fields='5-7'

#Remove spaces and get selected fields:
```

```
cat $input_file | tr -d ' ' | cut -d ',' -f $fields > tmp.txt
```

```
#Sort and get unique lines:
sort tmp.txt | uniq > $output_file
```

```
#Remove temporary file:
rm tmp.txt
```

También es posible introducir argumentos a un script, los cuales se llaman con \$1-9. \$0 representa el nombre del fichero.

```
#!/bin/bash
#Remove spaces and get selected fields:
cat $1 | tr -d ' ' | cut -d ',' -f $3 > tmp.txt

#Sort and get unique lines:
sort tmp.txt | uniq > $2

#Remove temporary file:
rm tmp.txt

#Show head of new file:
head -n 20 $2
```

Ejercicio: escribir un script que acepte tres argumentos: f, a, b. El argumento f es el nombre de un fichero csv, y los argumentos a y b son números integer. El script debe mostrar la lista de valores coincidentes en las columnas a y b del fichero f.

```
#!/bin/bash
file=$1
column1=$2
column2=$3

cut -d ',' -f $column1 $file | sort > tmp1.txt
cut -d ',' -f $column2 $file | sort > tmp2.txt
join -t ',' -1 1 -2 1 tmp1.txt tmp2.txt

rm tmp1.txt tmp2.txt
```

III.1.5. Editores de texto

vi es un editor de texto que viene en todas las distribuciones de Linux. La ventaja es que no abre ventanas externas (como gedit). Tiene dos modos: modo de inserción y modo de comando. En modo comando, los caracteres que se pulsan se interpretan como comandos, y se puede navegar mediante hjkl. Pulsando i, se cambia al modo de inserción, pudiendo añadir texto al editor. Para salir sin guardar, se debe poner :q!, y para salir guardando, :x. Para volver al modo de comando desde el modo inserción, basta con pulsar Esc.

III.2. Expresiones regulares y grep

El comando `grep` permite buscar texto, pero también patrones. Los patrones emplean distintos símbolos:

- `^` : inicio (por ejemplo, `^Hola`)
- `$` : final (por ejemplo, `Pepito$`)
- `.` : comodín, cualquier carácter (por ejemplo, `Pepit.` encuentra tanto Pepito como Pepita).
- `[]` : indican un conjunto de caracteres válidos (`Pepit[ao]` encuentra Pepito y Pepita, pero no Pepite).
- `[^]` : en la posición de los corchetes, vale cualquier carácter menos los que se especifiquen dentro (`Pepit[^o]` solo encuentra Pepita).
- `*` : indica un número indeterminado de repeticiones del carácter anterior, siendo 0 una posibilidad. Por ejemplo, con la expresión `"[1-9][0-9]*th"`, se encontrarían `01th`, `15th`, `2th` y `2333th`. De `01th`, solo se encontraría `1th` con el patrón, ya que el 0 no está incluido en el primer conjunto. Así, tanto `01th` como `2th` tienen el dígito en el primer corchete, el segundo corchete se repite 0 veces y finaliza en `th`. En el caso de `2333th`, el 2 se encontraría en el primer corchete, y los 3 en el segundo corchete, repitiéndose las veces necesarias.

Una expresión regular que extraiga las líneas que contengan solo números sería la siguiente:

```
#!/bin/bash
grep "^[0-9]*[0-9,]*$" $1
```

Para los símbolos especiales (como un punto o un circunflejo), se debe preceder el símbolo con una barra invertida para mostrar que el carácter es el que se quiere buscar, no por su significado en los patrones. Es decir, para buscar todas las líneas que empiecen con una letra mayúscula y terminen en un punto, se debe poner:

```
#!/bin/bash
grep '^ [A-Z] .*\. $'
```

Y las líneas que empiecen con circunflejo:

```
#!/bin/bash
grep '^ \^'
```

III.2.1. Expresiones reguladas extendidas

III.2.1.1. Agrupamientos

En algunas ocasiones se quieren agrupar expresiones. Esto se realiza con paréntesis, donde se especifica la cadena completa a buscar. Si queremos que los paréntesis se interpreten como delimitadores de grupos, se debe escapar de ellos con las barras invertidas.

```
#Buscar la cadena GTA seguida de cero o más repeticiones de GTA
grep 'GTA\ (GTA\)*'
#Buscar la cadena GTA seguida de cero o más repeticiones de (GTA)
grep 'GTA(GTA)*'
```

La opción -E permite interpretar patrones más extensos, entendiendo directamente los paréntesis como delimitadores de grupos. Así, cuando escapemos de ellos, lo que se interpreta es que se busque el símbolo del paréntesis dentro del patrón.

```
#Buscar la cadena GTA seguida de cero o más repeticiones de (GTA)
grep -E 'GTA\ (GTA\)*'
#Buscar la cadena GTA seguida de cero o más repeticiones de GTA
grep -E 'GTA(GTA)*'
```

III.2.1.2. Alternancia

También se pueden utilizar tuberías que funcionan como "or".

```
grep -E 'GACC|TCAG'
grep -E '(AA|GG|TT|CC)(AA|GG|TT|CC)*'
```

III.2.1.3. Cuantificadores

El símbolo ? indica una o ninguna repetición, mientras que + indica una o más repeticiones.

```
grep -E '(GACC)?' #0 o 1
grep -E '(GACC)+' #1 o más
grep -E '(GACC)*' #0 o más
```

También es posible especificar el número de veces que se repita una serie de caracteres:

```
#La cadena se repite exactamente 3 veces
grep -E '(GACC){3}'
#La cadena se repite entre 3 y 5 veces.
grep -E '(GACC){3,5}'
```

Ejercicios:

```
#Escribir una expresión regular que encuentre cadenas sin T y con un
  número impar de G
grep -E '[^TG]*G([GT]*G[GT]*G)*[GT]*'

#Escribir una expresión regular que encuentre direcciones de e-mail
grep -E '[a-z] (a-z0-9_)*\.(a-z0-9_)*@[a-z]*\.(es|com)'
```

III.3. Expansiones

III.3.1. Expansión aritmética

Para hacer operaciones aritméticas en la consola, se debe poner `$((operación))`.

```
echo $((4+3))
```

III.3.2. Expansión de llaves

Se pueden crear listas con llaves:

```
{list}

#Imprimir la lista A B C
echo {A, B, C}

#Imprimir la lista horse home hope horoscope
echo ho{rs, m, p, roscop}e
```

Se pueden utilizar dos puntos que implican el rango entre los dos valores. Así, `x..z` genera la lista `x` y `z`. También se puede anidar:

```
echo {z..a}
echo {1..9}
echo {X{1,2}, Y{3,4}}
echo {X,Y}{1..4}
```

III.4. Sustitución de comandos

Mediante la sustitución de comandos se puede utilizar la salida de un comando como expansión.

```
$(comando)
```

```
#Imprimir con echo la salida de ls
```

```
echo $(ls)
#Crear directorios con los nombres de los valores únicos de la cuarta
columna de adult.data
mkdir $(cut -d ',' -f 4 adult.data | sort | uniq)
```

III.5. El comando sed

Sed permite editar documentos de texto sin editor. Por ejemplo, sed 's/dog/cat/g' sustituye (s) cada aparición (g) de dog por cat. Esto lo saca por pantalla, pero se puede redireccionar a otro fichero. En el caso de sed '10,20s/dog/cat/g', la sustitución se realiza solo de las líneas 10 a 20. Si se eliminase la g, en una línea con varias apariciones solo se reemplazaría la primera.

Los comandos básicos son d para borrar la línea actual, p para imprimir la línea actual y n. La línea puede ser un número, un rango o una expresión regular.

```
sed '125s/dog/cat/' #sustituye solo la línea 125
sed '$s/dog/cat/' #sustituye la última línea del fichero
sed '1i\Header' #agrega el texto "Header" antes de la primera línea.
sed '2~2s/dog/cat/' #sustituye las líneas pares
sed '/^[A-Z]/s/dog/cat' #sustituye líneas que empiecen por mayúscula

sed '10,20d' #borra las líneas 10 a 20
sed '10,20!s/dog/cat/' #reemplaza dog por cat en todas las líneas
excepto de la 10 a la 20
```

También se pueden agrupar comandos.

```
sed '/^$/d' #elimina todas las líneas que empiezan y acaban sin tener
nada entre medias, es decir, las líneas en blanco
seq 1 20 | sed -n '3,10p' #imprime las filas 3 y 10 de la salida de seq
seq 1 20 | sed -n '/[26]/{s/1/5;p}' #muestra las líneas que tengan un 2
o un 6, sustituyendo en ellas el 1 por el 5
```

III.6. El comando awk

Este comando sirve para procesar texto línea a línea de cada fichero (como Excel en línea de comandos). awk sigue unas normas en la forma patrón {acción}.

```
awk 'BEGIN {print "Hello world"}' #imprime Hello world
awk '/regexp/ { print $0 }' input-file #imprime las líneas que cumplan
la expresión regular
awk '/regexp/' input-file #igual que antes; la acción predeterminada es
la impresión
awk '/regexp/ {}' input-file #no hace nada
awk 'length($0) > 50' input-file #imprime las líneas que tienen más de
50 caracteres
```

El ciclo awk es leer el fichero de entrada línea por línea. Por cada línea, awk intenta el patrón para cada norma. Para cada patrón que coincida, awk produce la acción de la norma de ese patrón. Si se cumplen varios patrones, entonces se ejecutan varias acciones. De igual forma, si no se cumple ningún patrón, no se ejecuta ninguna acción.

Awk lee la entrada en unidades llamadas registros. Los registros están separados por un carácter llamado separador de registros. El separador de registros es, por defecto, el carácter de nueva línea, por lo que los registros suelen ser líneas. El separador de registros puede cambiarse asignando un valor diferente a la variable RS. Awk lleva la cuenta del número de registros que se han leído en la variable NR.

Cada registro se divide en partes llamadas campos. Por defecto, los campos están separados por espacios en blanco (uno o más espacios, tabuladores o nuevas líneas). El separador de campos se puede cambiar asignando la variable FS, o en la línea de comandos con la opción -F. El número de campos de un registro se almacena en la variable NF. Se hace referencia a los campos con el signo \$:

- \$0 es todo el registro
- \$i, siendo i un número mayor que 0, es el campo número i.
- \$NF es el último campo

En el siguiente código, se imprime la longitud de la línea más larga. Al comienzo se inicializa la variable max que se va actualizando según se van pasando las líneas del fichero.

```
awk 'BEGIN { max = 0 }
    { if (length($0) > max) max = length($0) }
    END { print max }' adult.data
```

Otro ejemplo: imprimir la suma de los tamaños de los ficheros del usuario sandra:

```
ls -l | awk 'BEGIN { sum = 0 }
    $3 == "sandra" { sum += $5 }
    END {print sum }'
```

OFS es el separador de campo de salida. Así, se puede leer un fichero separado por comas y que el fichero de salida esté separado por punto y coma, por ejemplo.

```
#Se cambia el delimitador de espacio a dos puntos, el segundo valor (b)
  se reemplaza por un string vacío, se crea una sexta variable (nueva)
  con valor new, se imprime la línea y se imprime el número de
  registros.
```

```
echo a b c d | awk '{ OFS = ":"; $2 = "";
                    $6 = "new"; print $0;
                    print NF }'
```

```
#Se imprime el número total de registros (6, a-f), se cambia NF a un
  valor de 3 (a-c), se asigna a la variable 1 el valor de la variable 1
```

y se imprime toda la línea (que como se ha modificado NF a 3, la línea ya no es abcdef, si no solo abc).

```
echo a b c d e f | awk '{ print "NF =", NF; NF = 3;  
                        $1 = $1; print $0 }'
```

Capítulo IV

Bases de datos no relacionales

IV.1. Bases de datos no SQL

Una base de datos es un conjunto de datos interrelacionados. Los datos registran eventos significativos de cualquier tamaño y diferentes grados de complejidad. Para gestionar las bases de datos, hay sistemas de gestión de bases de datos (SGBD). Se trata de un software que permite *definir los datos a almacenar* (especificando el tipo de datos, su estructura y restricciones), *construir la base de datos para almacenar la información* y *manipular los datos* (consultar los datos, actualizarlos y generar informes). Por ejemplo, en el caso de una base de datos para la universidad, se tendría que especificar la estructura de los datos: 5 ficheros relativos a estudiantes, asignaturas, notas, requerimientos y grupos. Se debe especificar la estructura de cada una de las tablas y los tipos de datos para cada campo. Una vez hecho eso, se va guardando los datos en cada una de las tablas. Finalmente, se pueden manipular los datos para obtener los resultados a la lista de notas del estudiante Pérez, obtener los estudiantes registrados en la asignatura X, generar un nuevo grupo de la asignatura, etc. Las bases de datos relacionales se pueden representar gráficamente mediante modelos entidad-relación y modelos relacionales. La participación total se distingue de la parcial mediante puntas de flecha rellenas o vacías respectivamente. De igual forma, puede haber una o dos puntas de flecha para diferenciar las relaciones a uno y a muchos.

El sistema de gestión de bases de datos también debe asegurar la persistencia de datos, es decir, almacenar los datos de forma segura, mantener la consistencia de los datos (si alguna persona la modifica, que se actualice para todas las demás personas) y asegurar la disponibilidad de los datos mediante copias de seguridad.

IV.1.1. Motivación para las bases de datos no SQL

Las bases de datos no SQL tienen cuatro razones principales:

- **Escalabilidad:** se define como escalabilidad vertical el cambio de un ordenador o servidor a otro más potente. La desventaja es que se debe parar la base de datos puntualmente durante la migración del sistema. La alternativa es la escalabilidad horizontal, es decir, formar una red de varios ordenadores o servidores conectados.

- **Coste:** las bases de datos tradicionales funcionan mediante licencias basadas en el número de usuarios. En el caso de las aplicaciones web, no se puede estimar el número de usuarios que van a acceder a la base de datos a todas horas. Los sistemas de gestión de noSQL son de código abierto y, por tanto, gratuitos.
- **Flexibilidad:** las bases de datos relacionales dependen de un buen diseño de tablas y columnas desde el principio. Además, todas las filas de la tabla tienen las mismas columnas, lo que en algunos casos puede ser un desperdicio de espacio si muchos campos se deben mantener vacíos (por ejemplo, en el caso de comercios online que venden una gran variedad de productos, cada uno con características diferentes).
- **Disponibilidad**

Las bases de datos tradicionales o relacionales siguen un modelo ACID: atomicidad, consistencia, aislamiento y durabilidad. La atomicidad hace referencia a operaciones indivisibles, obligando a que se ejecuten todos los pasos o ninguno. La base de datos no se puede dejar en un estado en el que se viole la integridad de los datos; debe ser consistente. Además, las transacciones no son visibles a otros usuarios hasta que la ejecución se haya completado, siendo así operaciones aisladas. Finalmente, los datos deben estar guardados en un medio persistente para que tenga una gran durabilidad.

Las bases de datos no SQL siguen un modelo BASE: básicamente disponibles, estado suave (soft state) y eventualmente consistente. Como las bases de datos siempre están en una red de servidores, están protegidos de fallos parciales. Los servidores vecinos se pueden hacer cargo de las tareas de un servidor, permitiendo que los datos estén disponibles. En algunos casos, las bases de datos pueden estar en un estado inconsistente, pero entonces se activa el proceso de replicación. El estado suave hace referencia a que el estado puede cambiar a lo largo del tiempo, habiendo instantes en el que los datos se están replicando en los servidores.

IV.1.2. Not only SQL

NoSQL viene de Not Only SQL. La sintaxis utilizada es similar a SQL, pero no se utiliza como lenguaje de consulta. No se utilizan tablas ni join, y tienen una arquitectura distribuida (escalabilidad horizontal). Las ventajas son la flexibilidad de los datos. Se puede cambiar el esquema de las bases de datos sobre la marcha sin tener que parar la base de datos. Además, se pueden procesar en ordenadores con pocos recursos, y hay mecanismos de optimización para las consultas con grandes cantidades de datos.

Hay cuatro tipos de bases de datos noSQL:

- **Key-value:** estas bases de datos tienen claves que almacenan datos. Por ejemplo, en el comercio electrónico, se pueden utilizar varios pares clave-valor: accountNumber, Name, numItems y custType. Todos los clientes que utilicen esta web van a tener esas cuatro características, por lo que los nombres de las claves van precedidos de la clave primaria o un ID: cust1.accountNumber, cust2.name, cust1.numItems, etc para los customers y wrhs1.number wrhsw1.address, etc para los almacenes (warehouses). Lo

importante es que la clave no se repita; es una etiqueta única. Hay espacios de nombres (namespaces) que permite la colección de identificadores. Las claves deben ser únicas dentro de un espacio de nombres. Puede haber varios espacios de nombres dentro de una misma base de datos, denominados como bucket.

Los valores pueden ser tan simples como una cadena de texto o tan complejos como imágenes. Tienen una gran flexibilidad, ya que las cadenas pueden tener una longitud indeterminada, puede haber imágenes o no estar disponible. Es importante incluir comprobaciones explícitas en los programas, ya que hay libertad total a la hora de asignar cualquier valor a una clave.

En estas bases de datos no hay tablas, no hay columnas, no hay restricciones, no hay joins o claves externas y no se emplea SQL. No obstante, también hay algunas similitudes. Los buckets se pueden asemejar a los modelos relacionales.

- **Document-oriented:** los documentos hacen referencia a que la información almacenada está en formato JSON o XML. No es necesario crear un esquema previo, ya que al añadir un documento, las estructuras necesarias se crean. En las bases de datos noSQL de clave-valor, los documentos solo se podrían acceder mediante la clave, pero en este tipo se pueden utilizar lenguajes de consulta que permiten operaciones de comparación para buscar por un valor concreto. Se pueden almacenar documentos anidados, evitando así los join.
- **Column-oriented:** en estas bases de datos, una columna es un nombre con un valor asociado. Una fila sería un conjunto de columnas, pero diferentes filas pueden tener diferentes columnas. Cuando hay muchas columnas, se puede agrupar en colecciones. Estas bases de datos no requieren esquema y se pueden añadir columnas cuando se necesitan. Están pensadas para conjuntos de datos en los que hay muchas columnas. El lenguaje de consulta es similar a SQL, pero tiene una sintaxis específica.
- **Graph-oriented:** la forma de representación de los datos es el grafo. Cada nodo tiene flechas a objetos adyacentes. Es más rápido de buscar operaciones entre nodos que en las bases de datos relacionales, sobre todo en casos donde sería necesario utilizar join de forma recursiva. Estas bases de datos se suelen utilizar en redes sociales, sistemas de recomendación y manejo de redes.

Los grafos son un conjunto de nodos (vertex) relacionados mediante arcos (edges). Cuatro ejemplos de información que se puede representar en un grafo:

- **Localización geográfica:** muestra los caminos entre distintas localizaciones en ambos sentidos.
- **Enfermedades infecciosas:** los nodos van a ser personas, las cuales tienen propiedades (contagiado, inmune, no contagiado con enfermedad pasada, etc). Los arcos muestran la probabilidad de pasar la enfermedad.
- **Part-of:** relación de todos los componentes de vehículos
- **Redes sociales:** grafo bipartito con dos tipos de nodos (usuarios y post) y solo hay una relación (like de un usuario a un post).

Estas bases de datos representan una serie de ventajas:

Consultas más rápidas: En una base de datos tradicional, hay una tabla para estudiantes, asignaturas y matrícula. Para saber nombres de asignaturas en las que está matriculado cada estudiante, habrá que hacer un natural join de las tres tablas. En un grafo bipartito con nodos de estudiante y asignatura, encontrar las asignaturas de cada estudiante es más sencillo, ya que es simplemente seguir los caminos de cada nodo.

Modelaje más simple: En el caso de una red social, es más sencillo ver qué usuario da like a qué post que tener una tabla de usuarios, una de likes y una de posts y unirlos.

Múltiples relaciones entre nodos o entidades: Los grafos permiten facilitar la búsqueda de alternativas entre los distintos nodos.

Los grafos tienen cuatro elementos: nodos, arcos, caminos y ciclos. Los arcos pueden tener propiedades y estar o no dirigidos. En el programa Neo4j los grafos son todos dirigidos. Los caminos son la operación más frecuente, haciendo referencia a la sucesión de los nodos y arcos. Los ciclos se pueden interpretar como arcos de un nodo a sí mismo.

Los grafos tienen tres operaciones fundamentales. La *unión* de grafos muestra la conjunción de todos los nodos y arcos de los grafos. La *intersección* muestra los nodos y arcos en común. Las *operaciones de recorrido* de grafos permiten encontrar caminos que tengan determinadas especificaciones que se detallan en la consulta (empezar en el nodo X, terminar en el nodo Y, etc.).

Los grafos y nodos tienen varias propiedades:

- **Isomorfismos:** son dos grafos equivalentes cuando tienen los mismos nodos y arcos, aunque tengan disposición distinta.
- **Orden y tamaño:** el orden de un grafo es el número de nodos, mientras que el tamaño hace referencia al número de arcos. Un tipo de grafos especiales es el clique, el cual tiene nodos que están unidos todos con todos.
- **Grado:** el grado de un nodo es el número de arcos que inciden en el nodo. En general, da igual que entren o salgan los arcos. En el caso de que fuera dirigido, se habla de grado de entrada y de salida.
- **Cercanía:** muestra lo bien relacionado que está cada nodo, es decir, que con una menor cantidad de saltos entre nodos se llegue a la mayor cantidad de nodos. Esto es importante en las redes sociales. Se calcula mediante la fórmula $C(x) = \frac{1}{\sum_y d(y,x)}$.
- **Betweenness:** se puede definir como cuánto de cuello de botella es el nodo. En otras palabras, si el nodo se ve perjudicado, en qué medida afecta a los caminos entre los demás nodos.

Hay varios tipos de grafos:

- Grafos direccionados y no direccionados
- Multigrafos
- Bipartidos: tienen dos tipos de nodos con una sola relación.
- Grafos con peso o ponderados: los arcos tienen un valor, como la distancia entre ciudades.

- Redes de flujo: grafos en los que los arcos representan capacidad, como las emisiones de gas

IV.2. Pajek

Todas las representaciones de los grafos se encuentran en un fichero de extensión .net. Los elementos del grafo se escriben con un asterisco y se especifica la cantidad de elementos que tienen. Los vértices cuentan con un número y la etiqueta, que si se trata de una cadena de texto debe ir entre comillas. También pueden tener más etiquetas como la forma del nodo. Los edges muestran los identificadores de los dos nodos que se relacionan. También se pueden especificar distintos tipos de arcos con distintos colores. En la visualización, en caso de que alguna característica no se refleje, en Options se puede modificar. Además, se pueden generar subgrafos para cada tipo de arco desde la pestaña de Network en la pestaña principal y "Multiple Relations Network".

Los nodos pueden tener atributos, como nombre, sexo, lugar de trabajo y antigüedad. Esto se puede representar con la forma del nodo, el tamaño, etc. Esto se escribe en ficheros de extensión .vec que se deben cargar en Pajek en el apartado de vectores y seleccionar la casilla de la derecha. Se distinguen los edges (no dirigidos) de los arcs (dirigidos).