TALLER DE BIOCOMPUTACIÓN

ARN, Biopython y grafos.

Hugo Armando Guillén Ramírez

Grupo de Biocomputación, Ciencias de la Computación, CICESE.

 hguillen@cicese.edu.mx

 @HugoGuillen

 Hugo_A_Guillen_Ramirez

 /r/procrastinando

Octubre 2016

PREPARACIÓN

1. Descargar https://github.com/HugoGuillen/otono2016 | goo.gl/hJAuqk
2. Anaconda3

   CMD: conda install jupyter biopython networkx
3. ViennaRNA Package

   Añadir "C:\Program Files (x86)\ViennaRNA Package" al PATH
4. VARNA

   Verificar ejecución del applet
5. Jupyter

   CMD: jupyter notebook

CONTENIDO

1. Introducción
2. Plegamiento de ARN **(ViennaRNA Package 2, VARNA)**
3. Análisis de composición de secuencias **(Biopython)**
4. Dibujando redes de regulación **(NetworkX)**

INTRODUCCIÓN

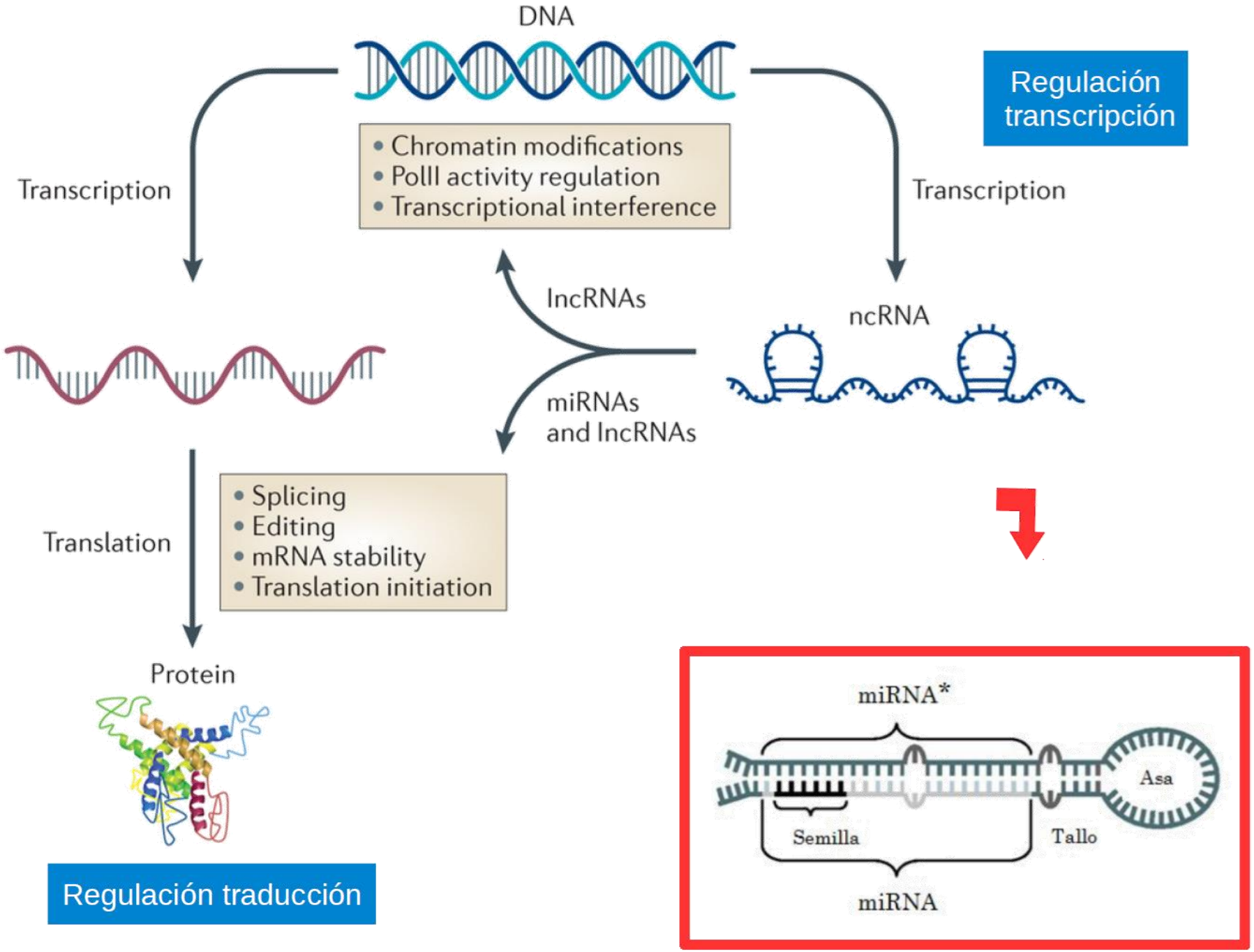


Imagen: ncRNA (gracias Diana Drago)

CLASIFICACIÓN

Etiquetar ejemplos **desconocidos** a partir de ejemplos **conocidos** (ya etiquetados).

Miniature Pinscher

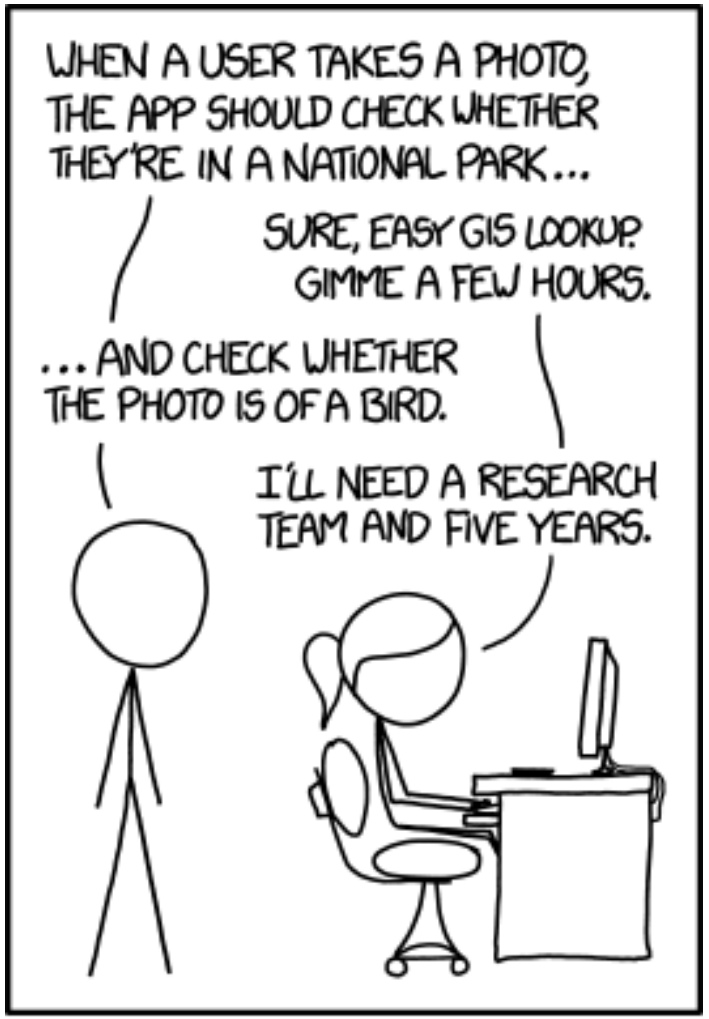
Miniature Schnauzer

Norfolk Terrier

Poodle
(Toy / Miniature)

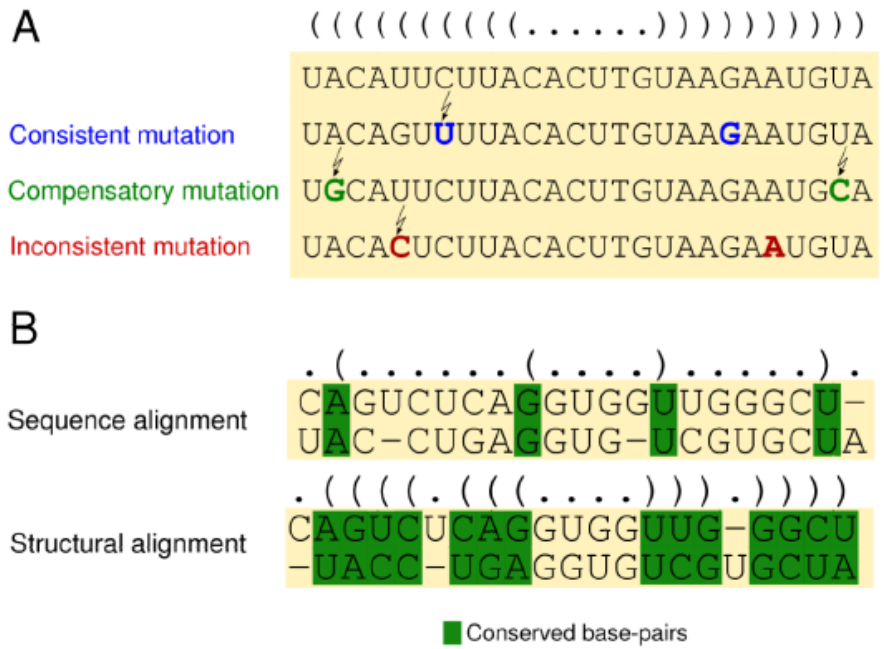
?

Schipperkee

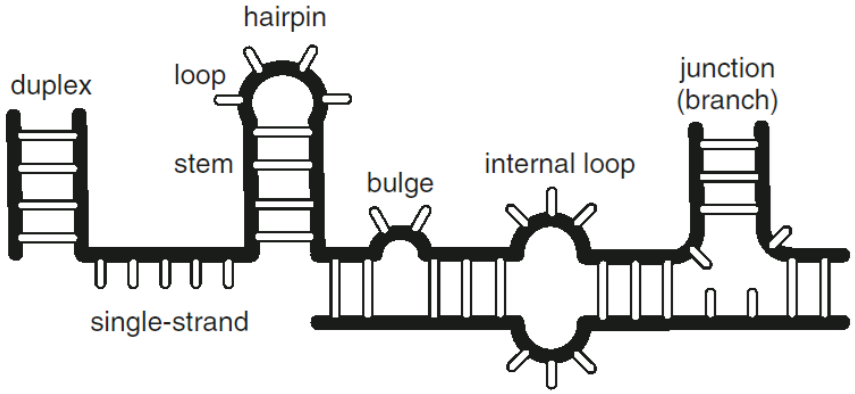


Características

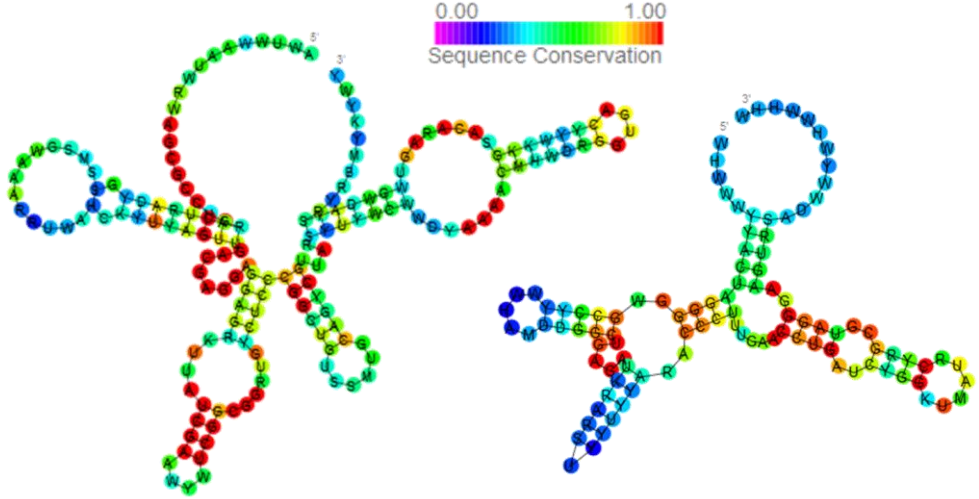
PLEGAMIENTO, ESTRUCTURA SECUNDARIA Y ALINEAMIENTOS



A) 3 secuencias que se pliegan en un hairpin alineadas.
B) Alineamientos a nivel de secuencia vs estructural.



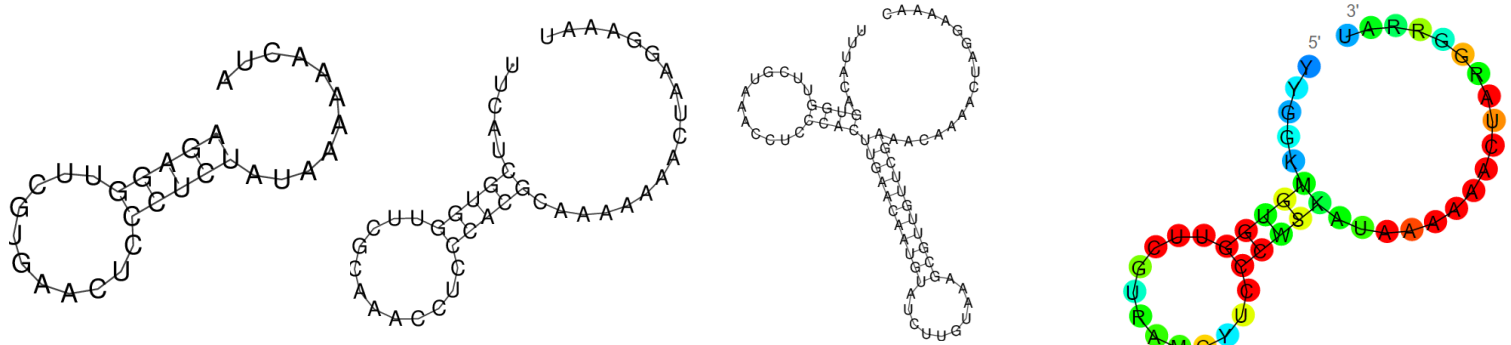
Nomenclatura de subestructuras de ARN



***glmS* (RF00234) [275 seqs]** **_TPP_ (RF00059) [3616 seqs]**

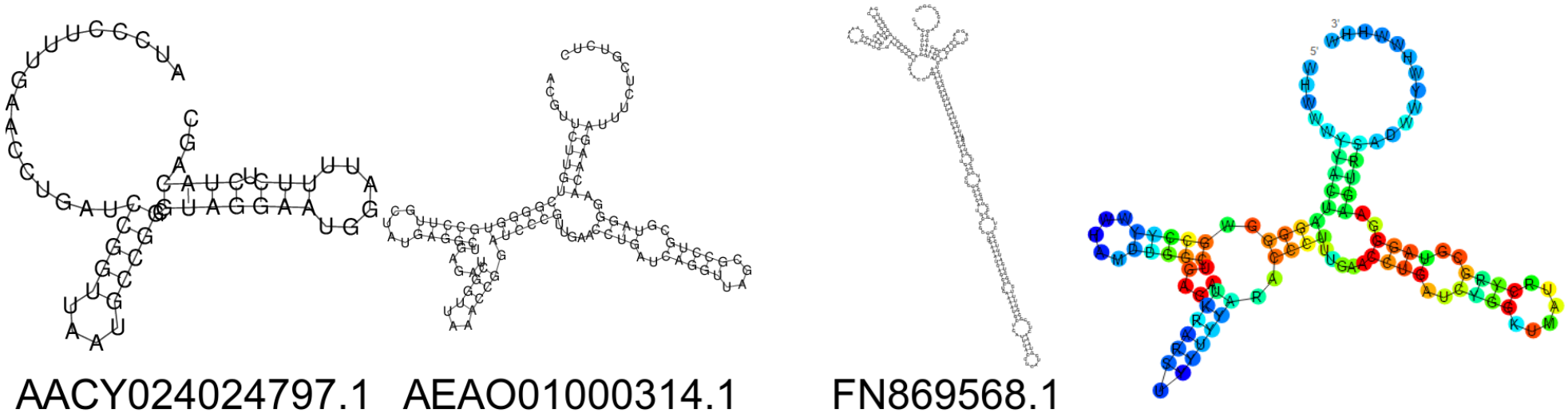
RFAM's consensus secondary structure of glmS and TPP.

Washietl, S., Will, S., Hendrix, D. A., Goff, L. A., Rinn, J. L., Berger, B., & Kellis, M. (2012). Computational analysis of noncoding RNAs. *Wiley Interdisciplinary Reviews: RNA*, 3(6), 759-778.



PreQ1

Estructura secundaria para secuencias de tamaño mínimo, medio, máximo, y estructura consenso RFAM de PreQ1



TPP

Estructura secundaria para secuencias de tamaño mínimo, medio, máximo, y estructura consenso RFAM de TPP

PLEGAMIENTO DE ARN

ViennaRNA Package 2

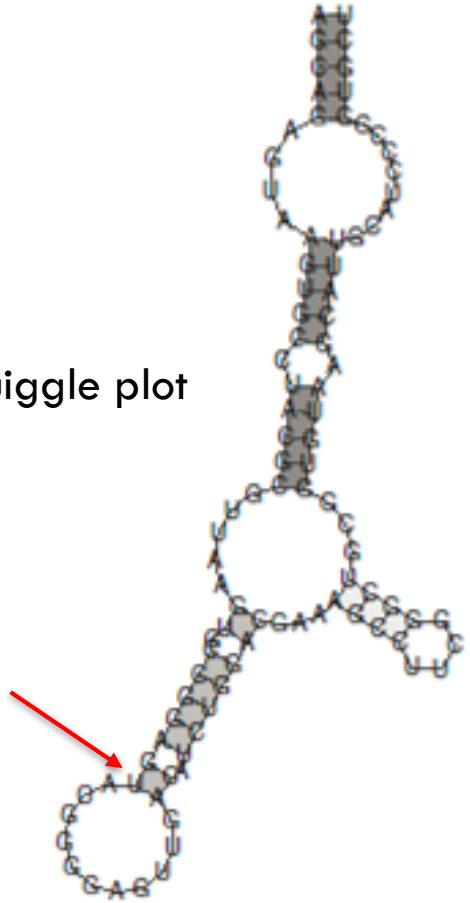
RNAfold

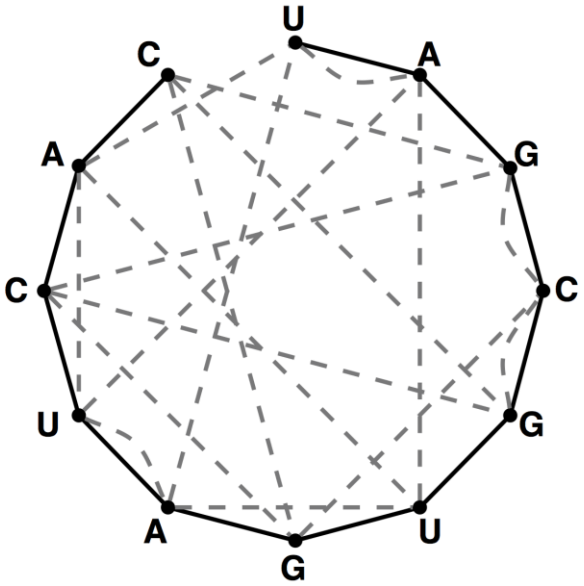
VARNA

RNA.ipynb

PLEGAMIENTO

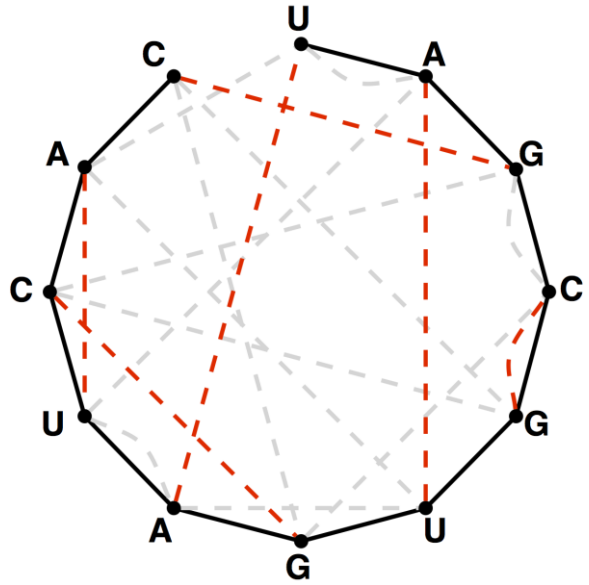
Squiggle plot





Bonding graph para UAGCGUGAUCAC.

Matching perfecto que representa una possible estructura secundaria.



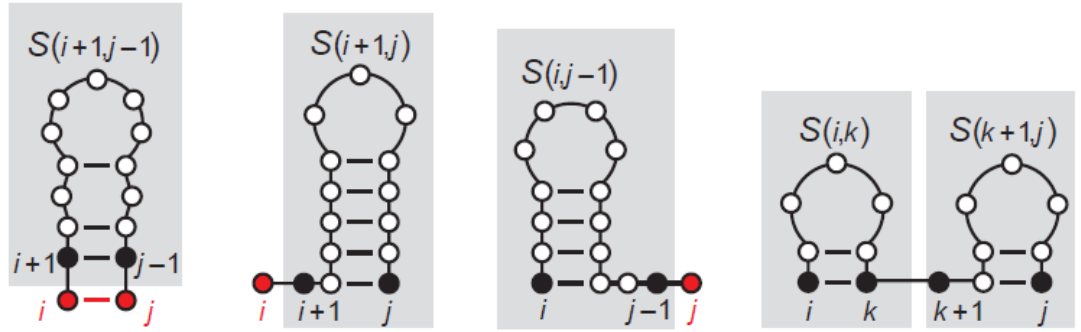

```
AGCAGAGUAAGUGCCUACGCGUUAAGUGCCGGAGUACGGGGAGUUGACAUCUGGACG
((((( .... ((((( .((((( ..... (( .((((((( .......... )) .)))))))
```

Notación dot-bracket

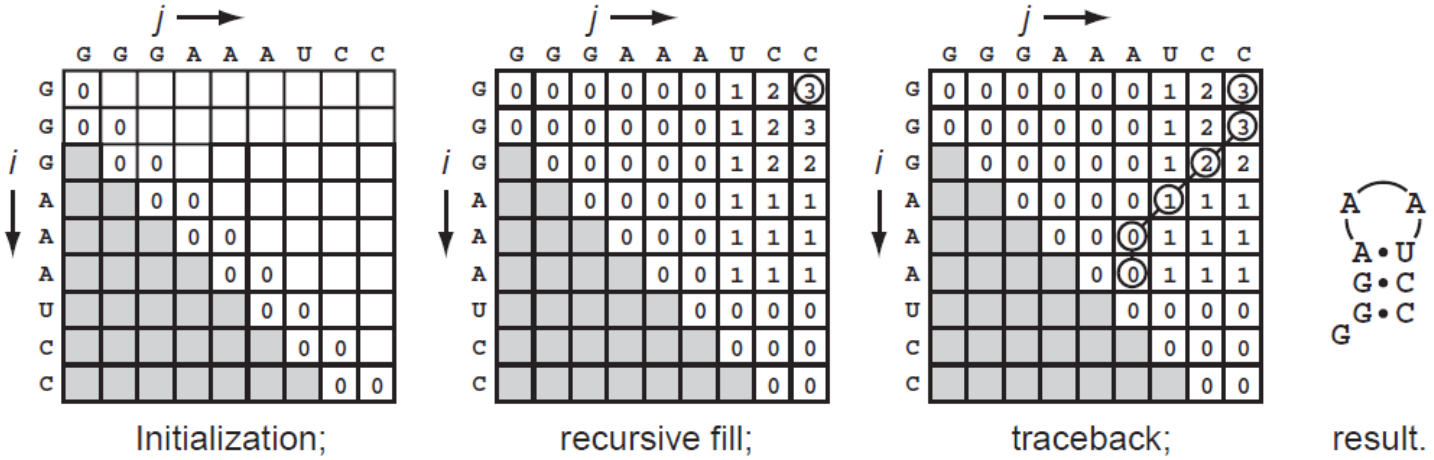
$$S(i,j) = \max \begin{cases} S(i+1,j-1)+1 & [\text{if } i,j \text{ base pair}] \\ S(i+1,j) \\ S(i,j-1) \\ \max_{i<k<j} S(i,k) + S(k+1,j) \end{cases}$$

Algoritmo de Nussinov y Jakobson (1980)

a Recursive definition of the best score for a sub-sequence *i,j* looks at four possibilities:



b Dynamic programming algorithm for all sub-sequences *i,j*, from smallest to largest:

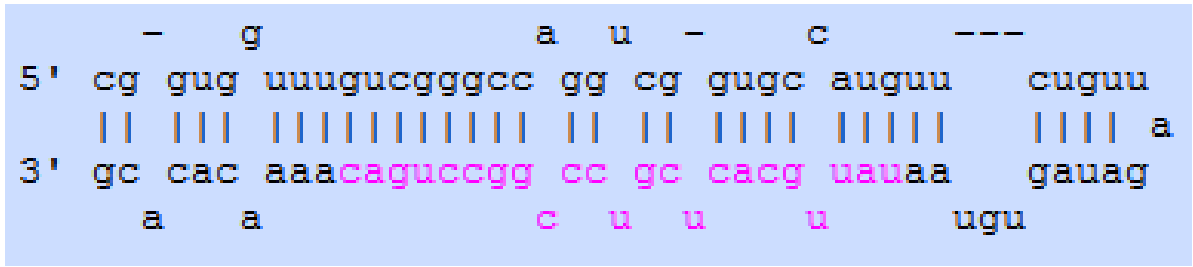


EJEMPLO: MIRNA

Xenoturbella bocki miR-92a stem-loop

The search for non invasive tools for diagnosis and management of cancer is extremely important for reducing the world wide health burden of cancer. miRNAs show potential as biomarkers and can even be found circulating in the serum. Some circulating miRNAs are specific to tumour patients, while miR-92 on the other-hand is present in healthy individuals in the serum but levels are variable and appear to change in response to the onset of some cancers.

>xbo-mir-92a MI0017684
CGGUGGUUUGUCGGGCCAGGUCGGUGCCAUGUUCUGUUAGAUAGUGUAAUAUUGCACUCGUCCCGGCCUGACAAAACACACG



http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0017684

RNAFOLD

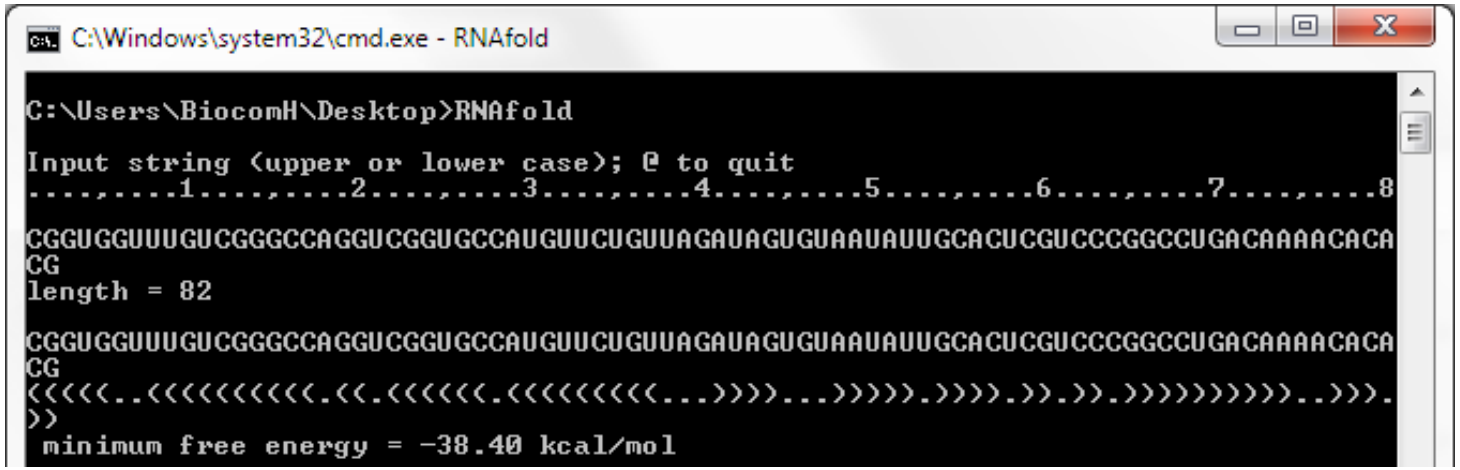
Calculate minimum free energy secondary structures and partition function of RNAs

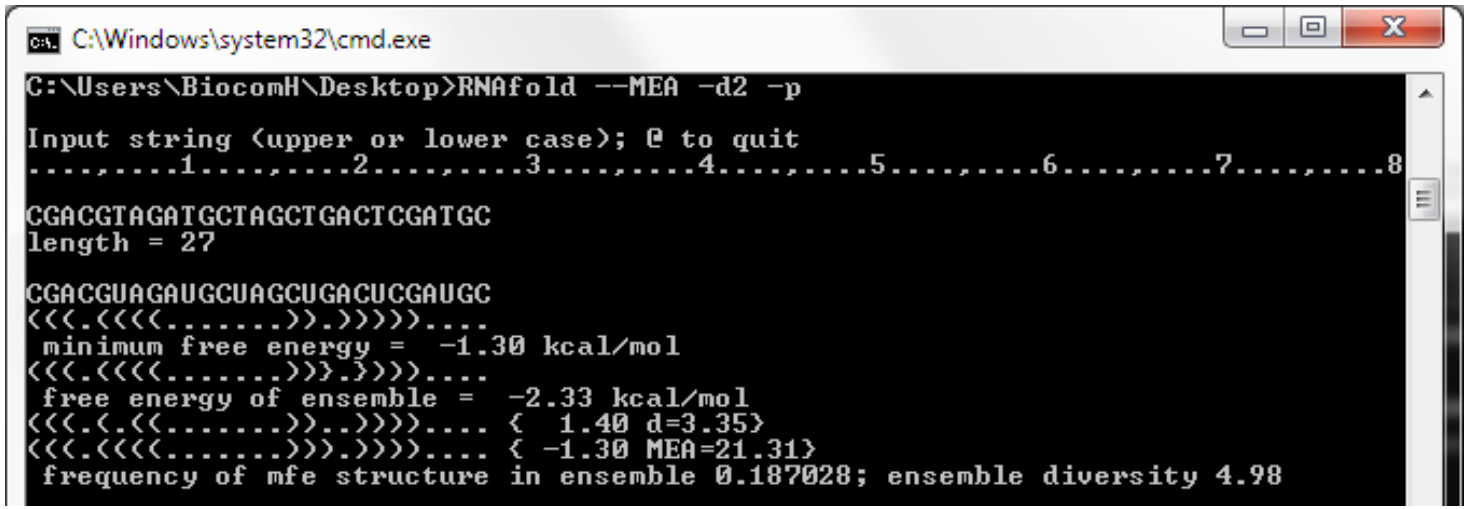
The program reads RNA sequences, calculates their minimum free energy (mfe) structure and prints the mfe structure in bracket notation and its free energy. If not specified differently using commandline arguments, input is accepted from stdin, and output printed to stdout. If the –p option was given it also computes the partition function (pf) and base pairing probability matrix, and prints the free energy of the thermodynamic ensemble, the frequency of the mfe structure in the ensemble, and the ensemble diversity to stdout.

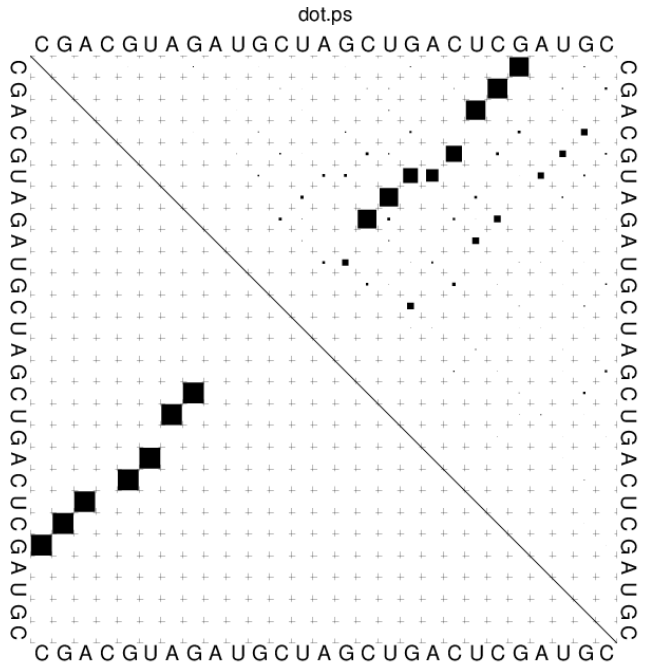
ACTIVIDAD 1

1. Ejecutar una consola en el Escritorio.
2. En la línea de comandos, ejecutar RNAfold.
3. Calcular estructura secundaria para la secuencia xbo-mir-92a MI0017684.
4. Ejecutar RNAfold --MEA -d2 –p
5. Calcular la estructura de la secuencia xbo-mir-92a MI0017684.

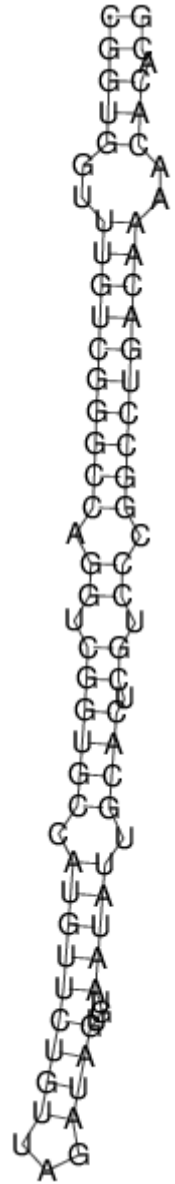
http://www.tbi.univie.ac.at/RNA/RNAfold.1.html


C:\Windows\system32\cmd.exe - RNAfold
C:\Users\BiocomH\Desktop>RNAfold
Input string (upper or lower case); @ to quit
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8
CGGUGGUUUGUCGGGCCAGGUCGGUGCCAUGUUCUGUUAGAUAGUGUAAUAUUGCACUCGUCCCGGCCUGACAAAACACA
CG
length = 82
CGGUGGUUUGUCGGGCCAGGUCGGUGCCAUGUUCUGUUAGAUAGUGUAAUAUUGCACUCGUCCCGGCCUGACAAAACACA
CG
(((((..((((((((((.((.((((((.(((((((((...))))...)))))).))))..))..))..)))))))))..)))).
))
minimum free energy = -38.40 kcal/mol


C:\Windows\system32\cmd.exe
C:\Users\BiocomH\Desktop>RNAfold --MEA -d2 -p
Input string (upper or lower case); @ to quit
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8
CGACGTAGATGCTAGCTGACTCGATGC
length = 27
CGACGUAGAUGCUAGCUGACUCGAUGC
(((.((((.......)).)))))....
minimum free energy = -1.30 kcal/mol
(((.((((.......))).))))....
free energy of ensemble = -2.33 kcal/mol
(((.(.((.......))..))))).... { 1.40 d=3.35}
(((.((((.......))).))))).... { -1.30 MEA=21.31}
frequency of mfe structure in ensemble 0.187028; ensemble diversity 4.98


dot.ps
CGACGUAGAUGCUAGCUGACUCGAUGC
CGACGUAGAUGCUAGCUGACUCGAUGC
CGACGUAGAUGCUAGCUGACUCGAUGC
CGACGUAGAUGCUAGCUGACUCGAUGC

dot.ps



rna.ps

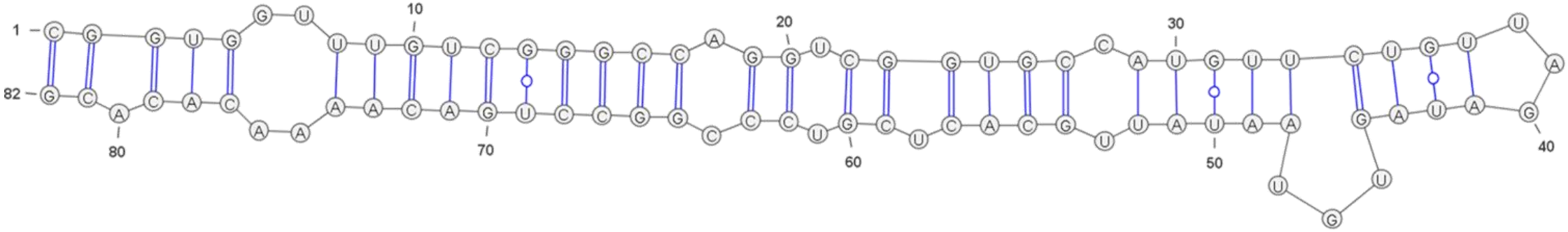
VARNA

¡Vamos a mejorar la visualización de la estructura secundaria!

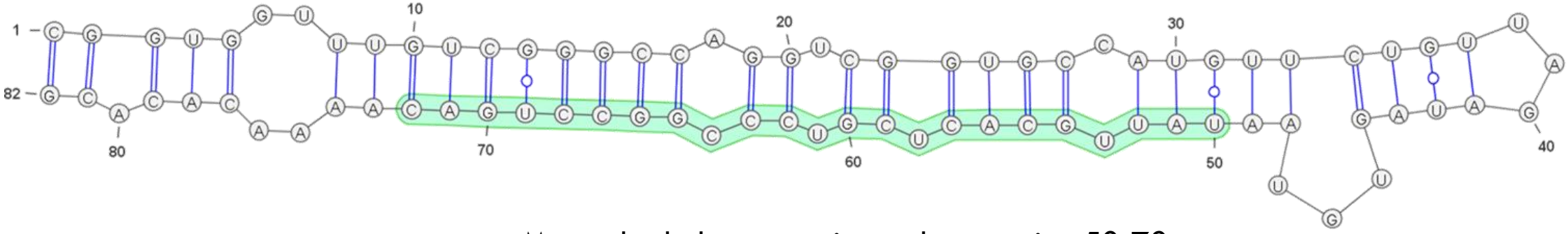
ACTIVIDAD 2

1. Ejecutar VARNA
2. Introducir secuencia y estructura generada por RNAFold en la Actividad 1.3
3. Encontrar un algoritmo que dibuje la estructura similar a la presentada en miRBase
4. La secuencia madura del miRNA está en la región 50-72. Marcar usando VARNA.

http://varna.lri.fr/



Visualización de xbo-mir-92a MI0017684 en VARNA usando el algoritmo NAView.



Marcado de la secuencia madura, region 50-72.

```
     -    g              a   u   -     c      ---
5'  cg  gug  uuugucgggcc  gg  cg  gugc  auguu    cuguu
    ||  |||  ||||||||||||  ||  ||  ||||  |||||    ||||  a
3'  gc  cac  aaacaguccgg  cc  gc  cacg  uauaa    gauag
      a    a              c   u   u     u       ugu
```

Representación de la estructura del miRNA en miRBase.

UNIÉNDOLO (ENVOLVIÉNDOLO) TODO CON PYTHON

`subprocess.` **`run`**(*args*, *, *stdin=None*, *input=None*, *stdout=None*, *stderr=None*, *shell=False*, *timeout=None*, *check=False*)

Run the command described by *args*. Wait for command to complete, then return a `CompletedProcess` instance.

`Popen.` **`communicate`**(*input=None*, *timeout=None*)

Interact with process: Send data to stdin. Read data from stdout and stderr, until end-of-file is reached. Wait for process to terminate. The optional *input* argument should be data to be sent to the child process, or `None`, if no data should be sent to the child. The type of *input* must be bytes or, if *universal_newlines* was `True`, a string.

`communicate()` returns a tuple `(stdout_data, stderr_data)`. The data will be bytes or, if *universal_newlines* was `True`, strings.

Cuando automatices…

1. Ejecuta el comando manualmente.
2. Verifica la llamada.
3. Revisa que se haya hecho bien.

https://docs.python.org/3/library/subprocess.html

LLAMANDO A RNAFOLD

ACTIVIDAD 3

Desarrollar una función en Python 3 con el siguiente comportamiento:

ENTRADA: una cadena de ADN/ARN

SALIDA: la estructura secundaria y la Mínima Energía Libre (MFE) asociada a la estructura secundaria de dicha cadena.

>hsa-let-7a-1 MI0000060
UGGGAUGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAUCUACUGUCUUUCCUA

http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000060

MARCANDO REGIONES DE INTERES EN ARN

ACTIVIDAD 4

Desarrollar una función en Python 3 con el siguiente comportamiento:

ENTRADA: una cadena de ADN/ARN, su estructura secundaria predicha, una región de interés, y la dirección donde VARNA escribirá la imagen de la estructura secundaria con la subsecuencia marcada.

SALIDA: ninguna.

java -cp VARNAvX-Y.jar fr.orsay.lri.varna.applications.VARNAcmd

[-i inputFile | -sequenceDBN XXX -structureDBN YYY] -o outFile [opts]

>hsa-let-7a-1 MI0000060
UGGGAUGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAUCUACUGUCUUUCCUA

>hsa-let-7a-5p MIMAT0000062
UGAGGUAGUAGGUUGUAUAGUU

http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000060

http://varna.lri.fr/index.php?lang=en&page=command&css=varna

ANÁLISIS DE COMPOSICIÓN DE SECUENCIAS

Biopython
Matplotlib
Biopython.ipynb

BIOPYTHON

Seq and SeqRecord objects

Bio.SeqIO - sequence input/output

Bio.AlignIO - alignment input/output

Bio.PopGen - population genetics

Bio.PDB - structural bioinformatics

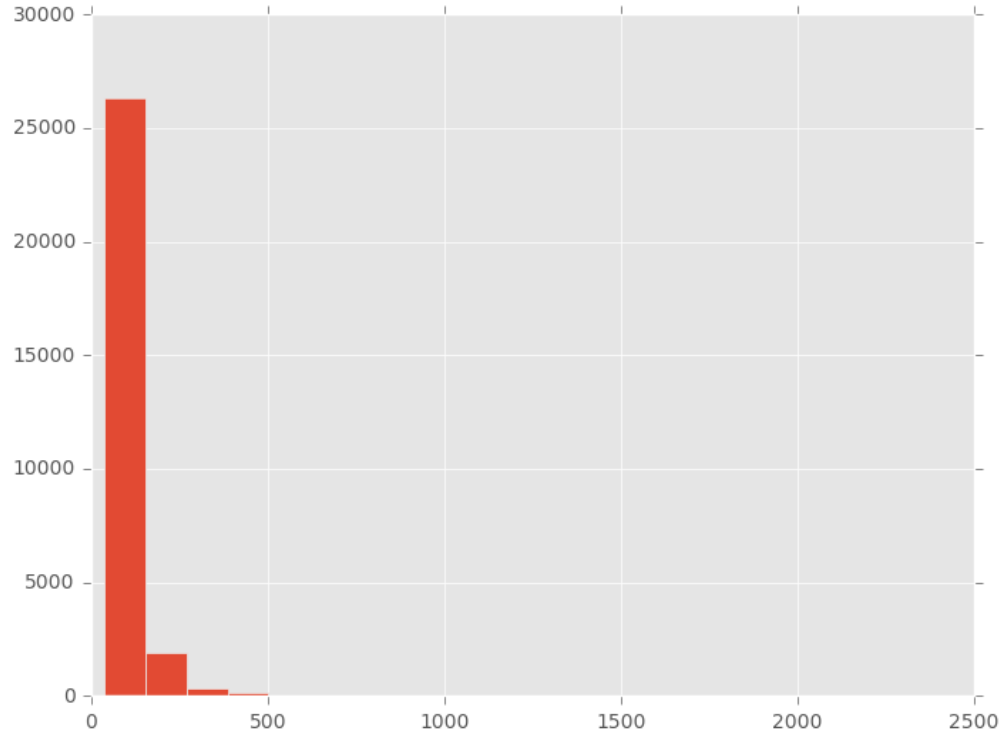
Biopython's BioSQL interface

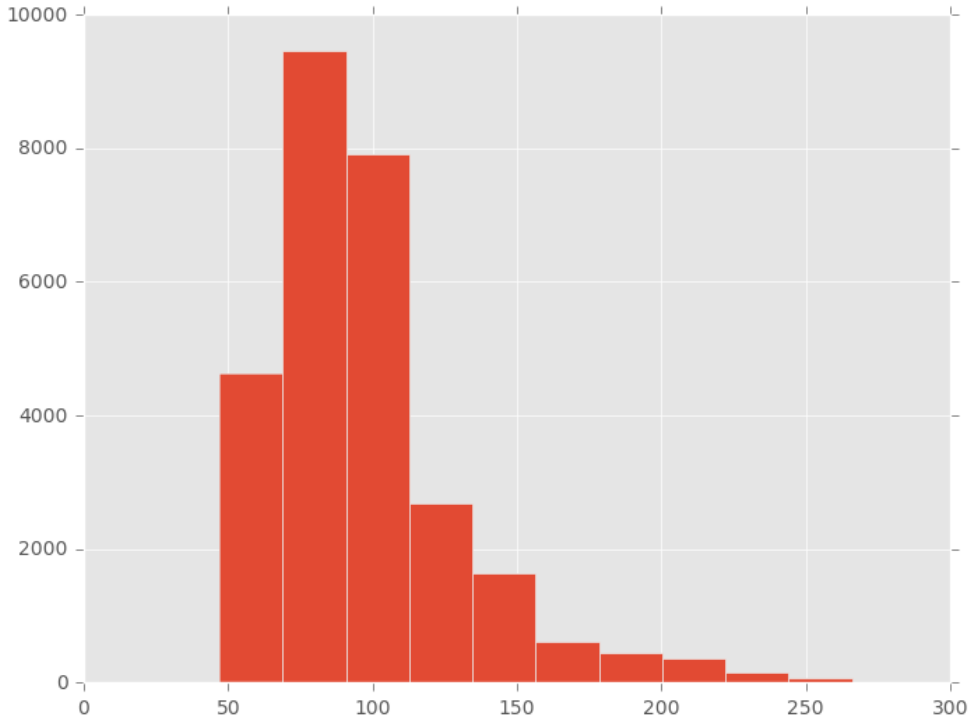
ACTIVIDAD 5

Desarrollar una función en Python 3 con el siguiente comportamiento:

ENTRADA: la dirección de un archivo fasta.

SALIDA: un histograma de las longitudes de las secuencias contenidas en el archivo


30000
25000
20000
15000
10000
5000
0
0
500
1000
1500
2000
2500

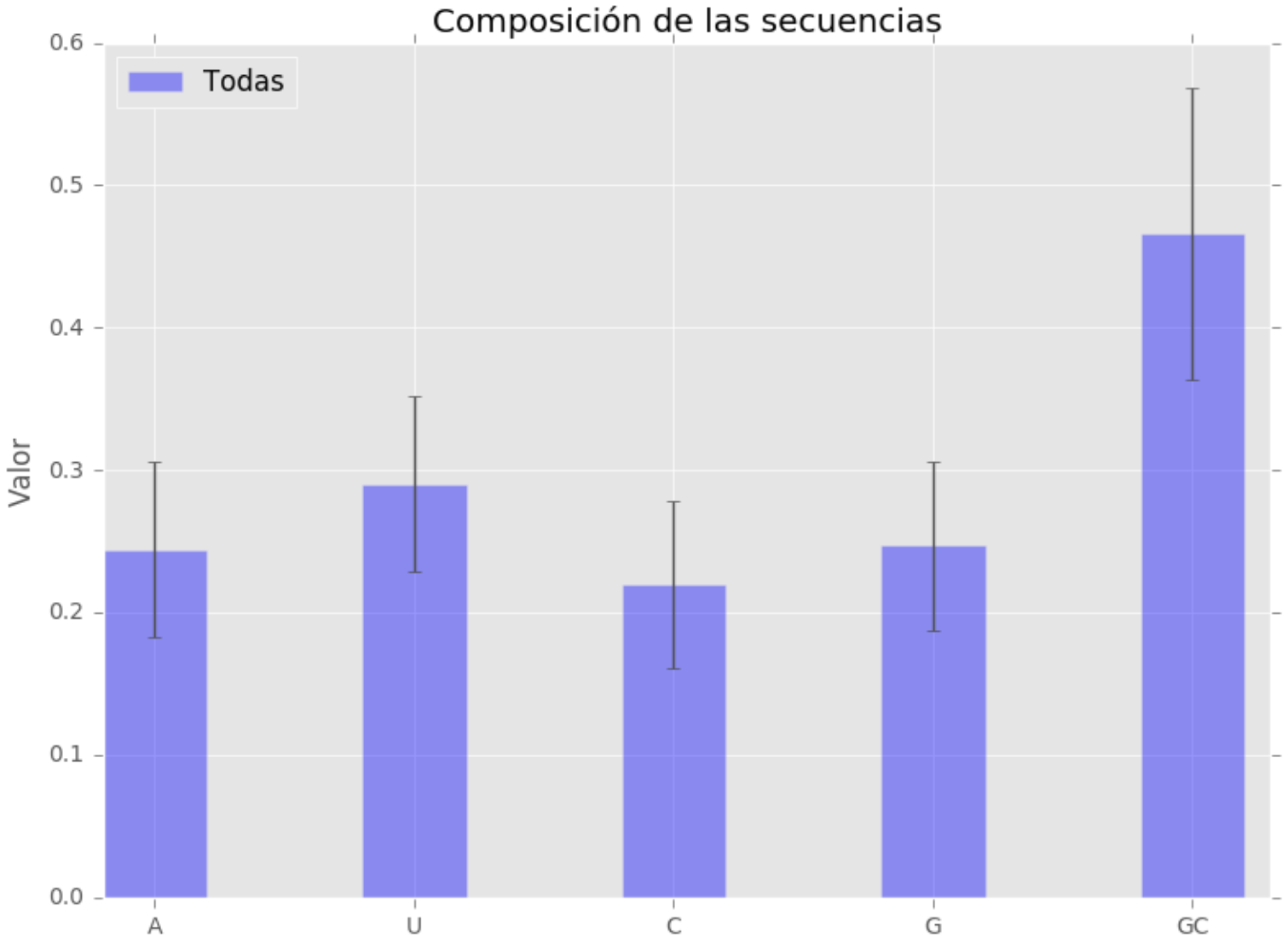

10000
8000
6000
4000
2000
0
0
50
100
150
200
250
300

ACTIVIDAD 6

Desarrollar una función en Python 3 con el siguiente comportamiento:

ENTRADA: la dirección de un archivo fasta.

SALIDA: un gráfico de barras que muestre la frecuencia de nucleótido y el contenido de CG.

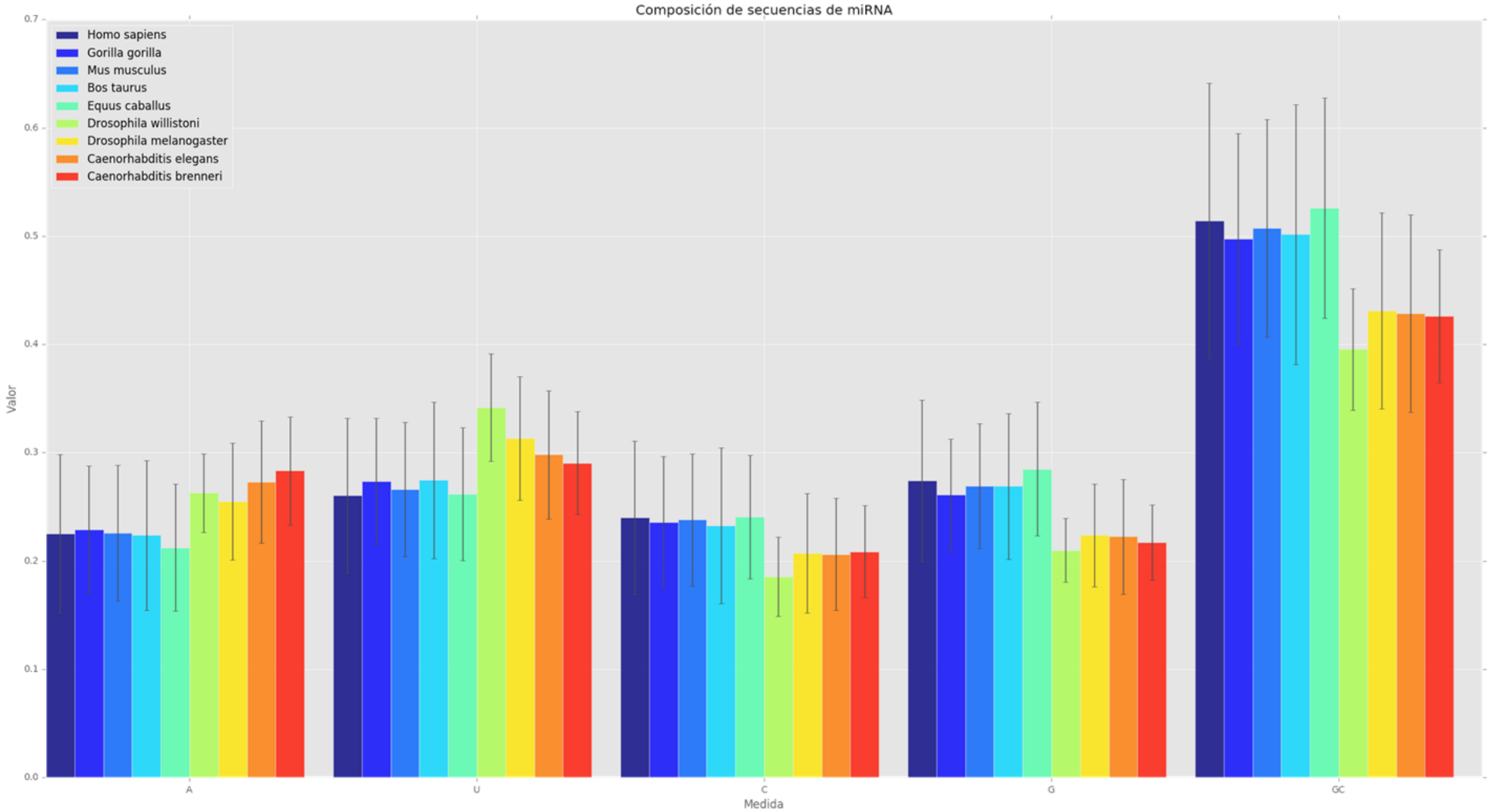

Composición de las secuencias
Todas
Valor
0.6
0.5
0.4
0.3
0.2
0.1
0.0
A
U
C
G
GC

ACTIVIDAD 7

Desarrollar una función en Python 3 con el siguiente comportamiento:

ENTRADA: el archivo fasta con las secuencias de hairpins de miRBase.

SALIDA: un gráfico de barras que muestre la frecuencia de nucleótido y el contenido de GC de las especies *'Homo sapiens', 'Gorilla gorilla', 'Mus musculus', 'Bos taurus', 'Equus caballus', 'Drosophila willistoni', 'Drosophila melanogaster', 'Caenorhabditis elegans‘ y 'Caenorhabditis brenneri'*


Composición de secuencias de miRNA
Homo sapiens
Gorilla gorilla
Mus musculus
Bos taurus
Equus caballus
Drosophila willistoni
Drosophila melanogaster
Caenorhabditis elegans
Caenorhabditis brenneri
Valor
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
A
U
C
G
GC
Medida

DIBUJANDO REDES DE REGULACIÓN

NetworkX

GRAPHS.ipynb

NETWORKX

High-productivity software for complex networks

NetworkX is a Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

Features

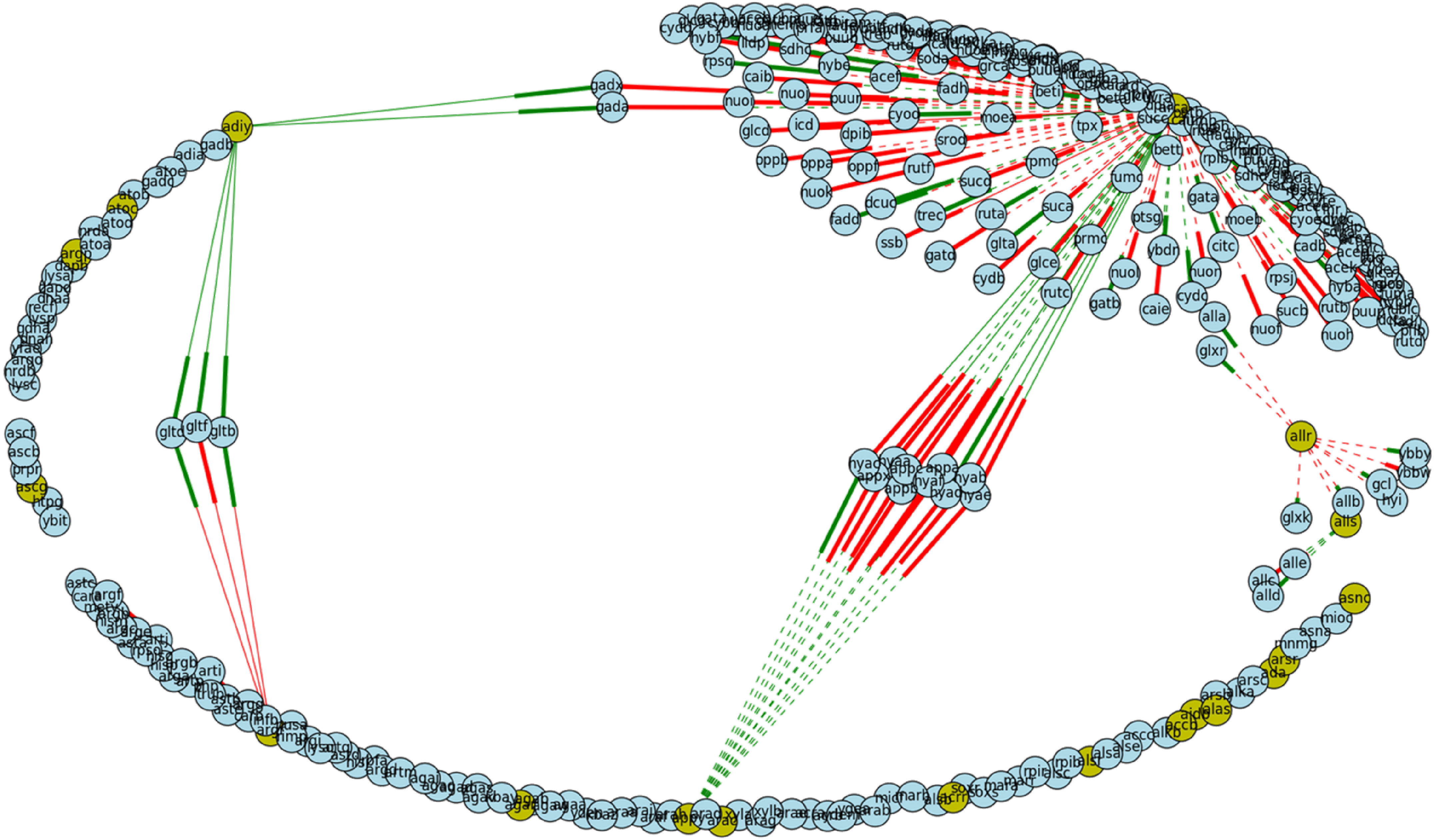
- Python language data structures for graphs, digraphs, and multigraphs.
- Many standard graph algorithms
- Network structure and analysis measures
- Generators for classic graphs, random graphs, and synthetic networks
- Nodes can be "anything" (e.g. text, images, XML records)
- Edges can hold arbitrary data (e.g. weights, time-series)
- Open source BSD license
- Well tested: more than 1800 unit tests, >90% code coverage
- Additional benefits from Python: fast prototyping, easy to teach, multi-platform

ACTIVIDAD 8

Desarrollar una función en Python 3 con el siguiente comportamiento:

ENTRADA: el archivo networw_tf_gene_mini.txt.

SALIDA: un grafo representativo de las relaciones entre los genes.



gadx
gada
adiy
gadb
adia
gltd
gltf
gltb
nuoi
caib
nuoj
puur
icd
dpib
glcd
oppb
oppa
oppf
rutf
nuok
dcuc
fadd
trec
ssb
gatd
cydb
glta
glce
rutc
prmc
suca
ruta
sucd
rpmc
fumc
tpx
moea
cyod
fadh
acef
hybe
sdhc
lldp
rpsg
beti
bett
ptsg
ybdn
nuol
gatb
caie
citc
moeb
rpsj
sucb
nuof
nuoh
rutb
alla
glxr
allr
ybby
ybbw
gcl
hyi
allb
alls
glxk
alle
allc
alld
asnc
mioc
asna
mnmg
arsc
ascf
ascb
prpr
ascg
htpg
ybit
astc
argf
arti
argb
xylb
arab
marb
alsa
alse

¡GRACIAS POR SU ATENCIÓN!