

Practica 3.

LAZARO CAMASCA EDSON NICK

1. Introducción

PRACTICA: ALINEAMIENTO CON BIOPYTHON Y CLUSTALO.

I. Leer Archivos de secuencias

P1. Describa los resultados obtenidos.

Al ejecutar el archivo readSqGBK.py que lee la secuencia en formato gbk nos proporciona información que esta tales como el código, Descripción como Protein X, Peptide X el nombre, y la secuencia misma.

```
nick@nick-VirtualBox: ~/Documentos/BiologiaComputacional/Secuencias
nick@nick-VirtualBox: ~/Documentos/BiologiaComputacional/Secuencias 120x15
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$ python readSqGBK.py
Codigo: P17102.1
Descripcion: RecName: Full=Protein X; AltName: Full=HBx; AltName: Full=Peptide X; AltName: Full=pX
Nombre: X_HBVA4
Secuencia: MATRLCCQLDPSRDVLCRLPVGAESRGRPLSGPLGTLSSPSPSAVPADHGAHLSLRGLPVCAFSSAGPCALRFTSARCMETTVNAHQILPKVLHKRTLGLPAMSTTDLE
AYFKDCVFKDWEELGEEIRLVFVLGGCRHKLVCAPAPCNFF TSA
```

P2. Describa los resultados obtenidos.

Al ejecutar el archivo readSqFasta.py que lee la secuencia en formato fasta nos proporciona informacion con una descripcion más específica como genotipo A2 del virus del Hepatitis B, además las secuencia misma.

```
nick@nick-VirtualBox: ~/Documentos/BiologiaComputacional/Secuencias
nick@nick-VirtualBox: ~/Documentos/BiologiaComputacional/Secuencias 120x15
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$ python readSqFasta.py
Descripcion: sp|P17102|X_HBVA4 Protein X OS=Hepatitis B virus genotype A2 subtype adw2 (isolate Germany/991/1990) GN=X P
E=3 SV=1
Secuencia: MATRLCCQLDPSRDVLCRLPVGAESRGRPLSGPLGTLSSPSPSAVPADHGAHLSLRGLPVCAFSSAGPCALRFTSARCMETTVNAHQILPKVLHKRTLGLPAMSTTDLE
AYFKDCVFKDWEELGEEIRLVFVLGGCRHKLVCAPAPCNFF TSA
```

2. Alineamiento de Secuencias

P3: ¿Cuántos registros tiene?

Al ejecutar el archivo python unirSq.py, reúne 44 registros

```
nick@nick-VirtualBox: ~/Documentos/BiologiaComputacional/Secuencias
python unirSq.py
Numero de registros: 44
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias
```

P4. ¿Cuál es el contenido de este archivo?

Se puede ver el nombre de las secuencias, sus nombres abreviados y la secuencia misma.

Además, se puede visualizar el script unirSq.py.

```

nick@nick-VirtualBox: ~/Documentos/BiologiaComputacional/Secuencias
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$ ls
HBV_10407.fasta      P17102.fasta      readSqGBK.py
LHBs                 P17102.gbk        unirsq.py
LHBs_variants.fasta  readSqFasta.py

nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$ cat LHBs_variants.fasta
>P12934 RecName: Full=Large envelope protein; AltName: Full=L gly
coprotein; AltName: Full=L-HBsAg; Short=LHB; AltName: Full=Large
S protein; AltName: Full=Large surface protein; AltName: Full=Maj
or surface antigen;
MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSHNPWDNFNPNKDHWEANQVGA
GAFGPGFTPPHGGLLGWSPQAQVLTTPVAPPPASTNRQSGRQPTIPSPPLRDSHPQAM
QWNSTTFHQALLDPRVRGLYFAPAGSSSGTVNVPVTTASPISSISRTGDPAPNMENITS
GFLGPLLVLAQAGFLLTRITIPQSLDSWNTSLNFLGGAPTCPGQNSQSPSNHSPTSCP
PICPGYRWMLRRFIIIFLLLLCIFLLVLLDQGMPLVCPCLPGSTTTSTGPKCTCTI
PAQGTSMFPSCCCTKPSDGNCTCIPISSWAFARFLWEGASVRFSLWLLVFPVQWVGL
SPTVWLSVIWMWYGNPSLYNLSPLFLLPIFFCLWVYI
>Q8JMY6 RecName: Full=Large envelope protein; AltName: Full=L gly
coprotein; AltName: Full=L-HBsAg; Short=LHB; AltName: Full=Large
S protein; AltName: Full=Large surface protein; AltName: Full=Maj
or surface antigen;
MGAPLSTARRGMQNLVSNPLGFFPDHQLDPLFRANSSSPDWFNTNKNWPMANKVGV
GGFGPGFTPPHGGLLGWSPQAQVLTTPVAPPPASTNRQSGRQPTIPSPPLRDSHPQAM
QWNSTTFHQALLDPRVRGLYFAPAGSSSETQNPAPTIASLTSSIFSKTGDPAMMENITS
GLLRPLLVAQVCFLLTKILTIPQSLDSWNTSLNFLGVPPGCPGQNSQSPISNHLPTSCP
PTCPGYRWMLRRFIIIFLLLLCIFLLVLLDQGMPLVCPCLPGSTTTSTGPKCTCTT
LAQGTSMFPSCCCTKPSDGNCTCIPISSWAFGKYLWEMASARFSLWLLVQFVQWCVGL
SPTVWLLVIWMWYGNPLCSILSPFILLPIFCYLWAST

nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias 67x36
GNU nano 2.8.6 Archivo: unirsq.py
from Bio import SeqIO
import os

records = []
for filename in os.listdir("LHBs"):
    handle = open("LHBs" + "/" + filename)
    record = SeqIO.read(handle, "swiss")
    records.append(record)

print "Numero de registros:", len(records)

SeqIO.write(records, "LHBs_variants.fasta", "fasta")

```

3. Realizar el alineamiento múltiple de secuencias utilizando Clustal

a) Utilizando la línea de comandos clustalw

```

nick@nick-VirtualBox: ~/Documentos/BiologiaComputacional/Secuencias 86x36
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$ cat LHBs_variants.
aln
CLUSTAL 2.1 multiple sequence alignment

P12934      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSHNPWDNFNPNKDHWEANQVGA
Q67867      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDRWPEANQVGA
P03140      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKQWPEANQVGA
Q76862      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWEANQVGA
P31869      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWEANQVGA
Q81162      -----MGTNLSVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKQWPEANQVGA
P31868      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWEANQVGV
Q9E654      MGGYSSKPRKGMGTNLVSNPLGLPDHQLDPAFGANSNNPDWDFNPNKDPWEANQVGA
Q998L9      MGGWSSKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWEANQVGA
Q913A6      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVRA
P03141      MGGWSSKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDDWPAANQVGV
Q91534      MGGWSSKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
P03142      -----MGTNLSVSNPLGLPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
P17101      MGGWSSKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
Q91C35      -----MGTNLSVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
Q4R1R8      MGGRLPKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
Q4R1S6      MGGWLPKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
P31873      MGGWSAKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWEANQVGV
Q02317      MGGWSSKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
Q998M2      -----MGQNLSTSNPLGFFPDHQLDPAFRANTNPDWDFNPNKDTWPDANKVGA

```

b) Utilizando el programa clustalo

```

nick@nick-VirtualBox: ~/Documentos/BiologiaComputacional/Secuencias 135x36
HBV_10407.fasta  LHBs_variants.aln  LHBs_variants.fasta  P17102.fasta  readSqFasta.py  unirsq.py
LHBs             LHBs_variants.dnd  LHBs_variants_o.aln  P17102.gbk   readSqGBK.py
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$ cat LHBs_variants_o.aln
CLUSTAL O(1.2.4) multiple sequence alignment

P12934      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSHNPWDNFNPNKDHWEANQVGA
Q8JMY6      MGAPLSTARRGMQNLVSNPLGFFPDHQLDPLFRANSSSPDWFNTNKNWPMANKVGV
P17397      -----MGTNLSVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
P31868      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWEANQVGV
P03140      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKQWPEANQVGA
Q05496      MGAPLSTARRGMQNLVSNPLGFFPDHQLDPLFRANSSSPDWFNTNKNWPMANKVGV
Q998M2      -----MGQNLSTSNPLGFFPDHQLDPAFRANTNPDWDFNPNKDTWPDANKVGA
P31873      MGGWSAKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWEANQVGV
Q998R4      MGAPLSTARRGMQNLVSNPLGFFPDHQLDPLFRANSSSPDWFNTNKNWPMANKVGV
P24025      -----MGQNLSTSNPLGFFPDHQLDPAFRANTNPDWDFNPNKDSWPDANKVGA
P17398      -----MGTNLSVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
Q69606      MGAPLSTARRGMQNLVSNPLGFFPDHQLDPLFRANSSSPDWFNTNKNWPMANKVGV
Q67875      -----MGQNLSTSNPLGFFPDHQLDPAFRANTNPDWDFNPNKDTWPDANKVGA
Q8JN07      MGAPLSTARRGMQNLVSNPLGFFPDHQLDPLFRANSSSPDWFNTNKNWPMANKVGV
P03141      MGGWSSKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDDWPAANQVGV
Q76862      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWEANQVGA
Q76926      MGGWSSKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
Q998L9      MGGWSSKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWEANQVGA
Q8JMY6      MGAPLSTARRGMQNLVSNPLGFFPDHQLDPLFRANSSSPDWFNTNKNWPMANKVGV
Q9QM10      -----MGQNLSTSNPLGFFPDHQLDPAFRANTNPDWDFNPNKDTWPDANKVGA
Q91813      -MGLSWTVPLEGKNLSASNPLGLPDHQLDPAFRANTNPDWDFNPNKDPWEANQVGV
P17101      MGGWSSKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
Q91534      MGGWSSKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
Q913A6      MGGWSSKPRQGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVRA
Q4R1R8      MGGRLPKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
Q9PW13      MGGWSSKPRKGMGTNLVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV
P17399      -----MGTNLSVSNPLGFFPDHQLDPAFGANSNNPDWDFNPNKDHWPANQVGV

```


P5. ¿Cuál es la longitud del alineamiento?

Al ejecutar el código **alineamiento.py**, se puede notar que la longitud el alineamiento es 400.

```
nick@nick-VirtualBox: ~/Documentos/BiologiaComputacional/Secuencias
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$ python alineamiento.py
Alignment lenght: 400
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$

GNU nano 2.8.6 Archivo: alineamiento.py
#Importamos el paquete AlignIO
from Bio import AlignIO

#creamos el objeto del alineamiento
alignment = AlignIO.read(open("LHBs_variants.aln"), "clustal")

#Imprimimos la longitud del alineamiento
print "Alignment lenght: ",alignment.get_alignment_length()
```

P6. ¿Cuál es el resultado? Haga un screenshot

Primero nos muestra información de la primera secuencia alignment[0].

Segundo la secuencia de las columnas 38, 39, 40

Tercero las secuencias de las 10 primeras columnas.

```
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$ python a
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$ python a
Alignment lenght: 400
ID: P12934
Name: <unknown name>
Description: P12934
Number of features: 0
Seq('MGGWSSKPRQGMGTNLSPNPLGFFPDHQLDPAFGANSHNPWDNFNPKDHWPE...VYI', SingleLe
tterAlphabet())
HNNNNNNNNNNNTNTNNNNNNAARARRRNEEEDEDESSSSSS
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNSSSSSS
PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
RESULTADO DE LAS 10 PRIMERAS COLUMNAS
columna 0 : MMMM-MMMMMM-M-MMM-----MM-MM-MMMMMM
columna 1 : GGGG-GGGGGG-G-GGG-----MMM-GG-GG-GGGGGG
columna 2 : GGGG-GGGGGG-G-GGG-----GGG-GG-GG-AAAAAA
columna 3 : WWWW-WYWWW-W-RWWW-----LLLL-WW-WW-PPPPPP
columna 4 : SSSS-SSSSS-S-LLSS-----SSSS-SS-SS-LLLLLLL
columna 5 : SSSS-SSSSS-S-PPAS-----WWW-SS-SS-SSSSSS
columna 6 : KKKK-KKKKK-K-KKK-----TTT-KK-KK-TTTTTT
columna 7 : PPPP-PHPPP-P-PPP-----VVV-PP-PP-AAATTT
columna 8 : RRRR-RRRRR-R-RRR-----PPP-RR-RR-RRRRRR
columna 9 : QQQQ-QKKQK-K-KKK-----LLL-KK-KK-RRRRRR
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$

GNU nano 2.8.6 Archivo: alineamiento.py
#Importamos el paquete AlignIO
from Bio import AlignIO

#creamos el objeto del alineamiento
alignment = AlignIO.read(open("LHBs_variants.aln"), "clustal")

#Imprimimos la longitud del alineamiento
print "Alignment lenght: ",alignment.get_alignment_length()

print alignment[0]
#print alignment.get_column(38)
#print alignment.get_column(39)
#print alignment.get_column(40)

#Columna 38
print alignment[:, 38]
print alignment[:, 39]
print alignment[:, 40]

print "RESULTADO DE LAS 10 PRIMERAS COLUMNAS"
for i in range(10):
    print "columna ",i," : ",alignment[:,i]
```

3. Conseguir la información sumariada resultante:

P7. ¿Cuál es la secuencia de consenso?

```
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$ python sumary.py
Alignment lenght: 400
MGGXSSXXRXGMXNLSVNPPLGFFPDHQLDPAFXANSXNPWDNFNPKDXWPXANXVGXGAFGPGF
TPPHGGLLWSPQAQGLTTXPAPPPASTNRQSGRQPTXPSPPLRDXHPQAMQNSTXFHQXLXDP
RVRGLYFPAGGSSSGTVNPXPXASXISXISXSTGDPAXNMENITSGXLGPLLVLAGFLLTXILT
IPQSLDSMWTSLNFLGGXPXCXGQNSQSPSTNSHPTSCPPXCPGYRWMCLRRFIIFLLCLIFL
LVLLDYQGMPLVCPCLXPXSXTTSTGPCXTCTTAXAGTSMFPSCCCTKPDGNCCTCIPSSWAFXXK
LHEWASXRFSLVLLVPFVQWVGLSPTVNLVIMMMYWGPSLYXILSPFXPLPIFFCLWVYI
nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$

nick@nick-VirtualBox:~/Documentos/BiologiaComputacional/Secuencias$
GNU nano 2.8.6 Archivo: sumary.py
from Bio.Align import AlignInfo
from Bio import AlignIO

#creamos el objeto del alineamiento
alignment = AlignIO.read(open("LHBs_variants.aln"), "clustal")

#Imprimimos la longitud del alineamiento
print "Alignment lenght: ",alignment.get_alignment_length()

#objeto para informacion sumariada
summary_align = AlignInfo.SummaryInfo(alignment)

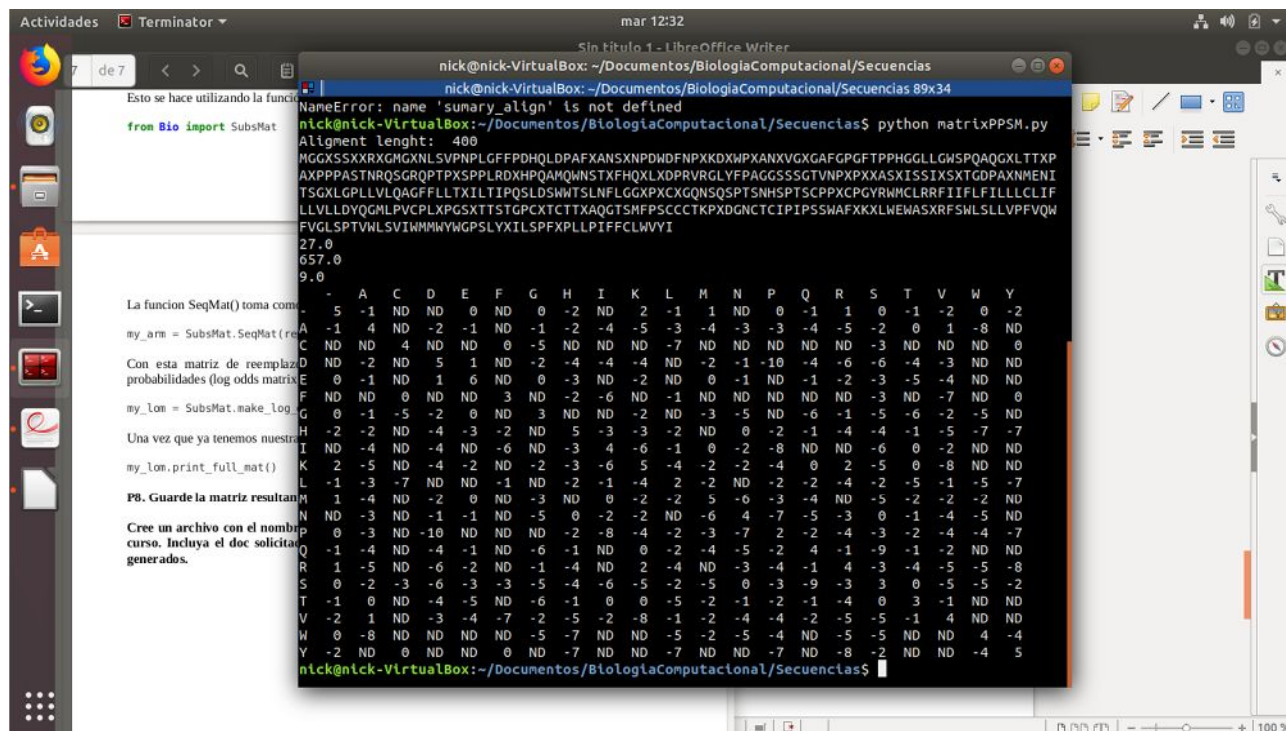
#calcular una secuencia de consenso simple
consensus = summary_align.dumb_consensus()

print consensus
```

4. Matrices de Score de Posiciones Especificas (PSSMs).

P8. Guarde la matriz resultante y agréguela al archivo de reporte del lab con sus respuestas.

Al ejecutar el script matrixPPSM.py obtenemos la matriz log odds



The screenshot shows a terminal window titled "nick@nick-VirtualBox: ~/Documentos/BiologiaComputacional/Secuencias" with a timestamp of "mar 12:32". The terminal displays the execution of a Python script named "matrixPPSM.py". The script outputs the alignment length (400) and a sequence of amino acids: "MGGXSSXXRXMGXNLVNPPLGFFPDHQLDPAFXANSXNPDPDFNPXKDXMPXANXVGXGAFGPGFTPPHGGLLGWSPQAQGXLTTPXAXPPASTNRQSGRQPTXSPPLRDXHPQAMQWNTXFXHQLXDPVRVGLYFPAGGSSSGTVNPPXXASXISSIXSXTGDPAXNMENITSGXLGPLLVLQAGFFLLTXILTIQSLDSWNTSLNFLGGXPXCXGQNSQSPTSNHSPTSCPPXCPGYRWMLRRFIIFLFIILLCLIFLLVLLDYQGMPLVCPXPGSXTTSGPCCTCTTQAQGTSMFPPSCCTKPDGNCCTCIPSSWAFKXKLWEWASXRFSLVLLVPFVQW". The script then outputs the log odds matrix, which is a 20x20 matrix with columns labeled A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. The matrix values are as follows:

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
A	-5	-1	ND	ND	0	ND	0	-2	ND	2	-1	1	ND	0	-1	1	0	-1	-2	0	-2
C	-1	4	ND	-2	-1	ND	-1	-2	-4	-5	-3	-4	-3	-3	-4	-5	-2	0	1	-8	ND
D	ND	ND	4	ND	ND	0	-5	ND	ND	ND	-7	ND	ND	ND	ND	-3	ND	ND	ND	0	
E	ND	-2	ND	5	1	ND	-2	-4	-4	ND	-2	-1	-10	-4	-6	-6	-4	-3	ND	ND	
F	0	-1	ND	1	6	ND	0	-3	ND	-2	ND	0	-1	ND	-1	-2	-3	-5	-4	ND	ND
G	ND	ND	0	ND	ND	3	ND	-2	-6	ND	-1	ND	ND	ND	ND	-3	ND	-7	ND	0	
H	0	-1	-5	-2	0	ND	3	ND	ND	-2	ND	-3	-5	ND	-6	-1	-5	-6	-2	-5	ND
I	-2	-2	ND	-4	-3	-2	ND	5	-3	-3	-2	ND	0	-2	-1	-4	-4	-1	-5	-7	-7
K	ND	-4	ND	-4	ND	-6	ND	-3	4	-6	-1	0	-2	-8	ND	ND	-6	0	-2	ND	ND
L	2	-5	ND	-4	-2	ND	-2	-3	-6	5	-4	-2	-2	-4	0	2	-5	0	-8	ND	ND
M	-1	-3	-7	ND	ND	-1	ND	-2	-1	-4	2	-2	ND	-2	-4	-2	-5	-1	-5	-7	
N	1	-4	ND	-2	0	ND	-3	ND	0	-2	-2	5	-6	-3	-4	ND	-5	-2	-2	-2	ND
P	ND	-3	ND	-1	-1	ND	-5	0	-2	-2	ND	-6	4	-7	-5	-3	0	-1	-4	-5	ND
Q	0	-3	ND	-10	ND	ND	ND	-2	-8	-4	-2	-3	-7	2	-2	-4	-3	-2	-4	-4	-7
R	-1	-4	ND	-4	-1	ND	-6	-1	ND	0	-2	-4	-5	-2	4	-1	-9	-1	-2	ND	ND
S	1	-5	ND	-6	-2	ND	-1	-4	ND	2	-4	ND	-3	-4	-1	4	-3	-4	-5	-5	-8
T	0	-2	-3	-6	-3	-3	-5	-4	-6	-5	-2	-5	0	-3	-9	-3	3	0	-5	-5	-2
V	-1	0	ND	-4	-5	ND	-6	-1	0	0	-5	-2	-1	-2	-1	-4	0	3	-1	ND	ND
W	-2	1	ND	-3	-4	-7	-2	-5	-2	-8	-1	-2	-4	-4	-2	-5	-5	-1	4	ND	ND
X	0	-8	ND	ND	ND	-5	-7	ND	ND	-5	-2	-5	-4	ND	-5	-5	ND	ND	4	-4	
Y	-2	ND	0	ND	ND	0	ND	-7	ND	ND	-7	ND	ND	-7	ND	-8	-2	ND	-4	5	