



Introducción a la línea de comando y a la programación para análisis bioinformáticos.



Teórico 3

Comandos para el manejo de textos.

Porque textos?

- La mayoría de los archivos que manejamos en bioinformática son de texto plano.

FASTA:

>aaf.1833 | aligned:- (160)

MSNQENIDITQLYKDGWIDERTGLMWSRPSLGQQWQKGKAIGSAEEFDFFQAQQTATTFTLAGFNDWQVPTLDELKTLLRDK
KKAGYDCPKGALFRPIEGSTWWGWFWSSTGGGTYSQIVAFTDGTANFTFSENKNYVRLVRHCPDTLKTQFIHASEN*

>aff.673 | aligned:- (160)

MSNQENIDITQLYKDGWIDERTGLMWSRPSLGQQWQKGKAIGSAEEFDFFQAQQTATTFTLAGFNDWQFFPTLDELKTLLRD
KKKAGYDCPKGALFRPIEGSTWWGWFFSSTGGGTYSQIVAFTDGTANFTFSENKNYVRLVRHCPDTLKTQFIHASEN*

FASTQ:

@071112_SLXA-EAS1_s_7:5:1:817:345
GGGTGATGGCCGCTGCCGATGGCGTCAATCCCACC
+071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC

SAM:

@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1



CLUSTAL W (1.83) multiple sequence alignment

abx.2190	MATIQKNLEMVNFVCKFGDADLLDLFDEVVYPAFFNLNNKRKFRGSEYYFKNVELLDLGK
cas.2954	MATIQKNLEMVNFVCKFGDADLLDLFDEVVYPAFFNLNNKRKFRGSEYYFKNVELLDLGK
dex.897	LNDTHMATLEKNLEIANLVCKFGDHNLLDLYKEVVEPSFFKNKNRKAGKSNFYFDEVSL

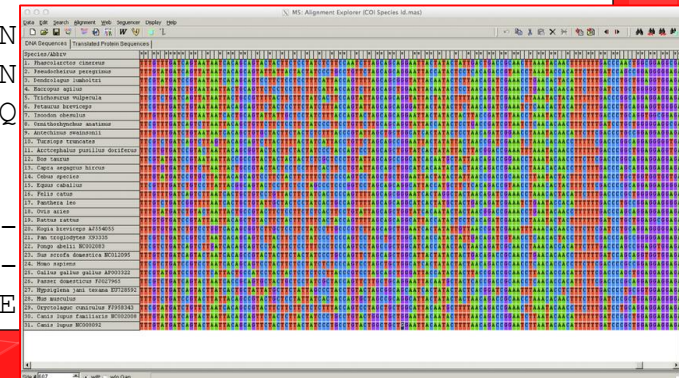
abx.2190	DPLGTKNLIAIAGRFIKNTLSREQYENDKGELIRDTQGMPSAPSAIFVLMLNSHRLLYV
cas.2954	DPLGTKNLIAIAGRFIKNTLSREQYENDKGELIRDTQGMPSAPSAIFVLMLNSHRLLYV
dex.897	VNLGDDILGEKDTLALVGRFVKQTKLEREQYRAATGKLVKDKKKLDSAPSAIFALILNNH

abx.2190	AETKFAPNLNTFKATVDSFFRSEKNKYNKKLDRDEQAKFKVRFSGSPEITITPLSTEKNVF
cas.2954	AETKFAPNLNTFKATVDSFFRSEKNKYNKKLDRDEQAKFKVRFSGSPEITITPLSTEKNVF
dex.897	RLLYAAETKYAPT PENFRSTLSYFLRNEQIN YKDSL PKEERNPFLKEFP IVELTLTPLTN

abx.2190	DHLRQFEKIKKVKVTLNDRNSEFNASKMFDDIHLATKKMNSTTSVEFKNLKEGLDKEETP
cas.2954	DHLRQFEKIKKVKVTLNDRNSEFNASKMFDDIHLATKKMNSTTSVEFKNLKEGLDKEETP
dex.897	QKTLSENIYLFNKLKKVTFILNDRNSEIDGEDMFRQMQETSDEAGGQKGFEITNNEGLNQ

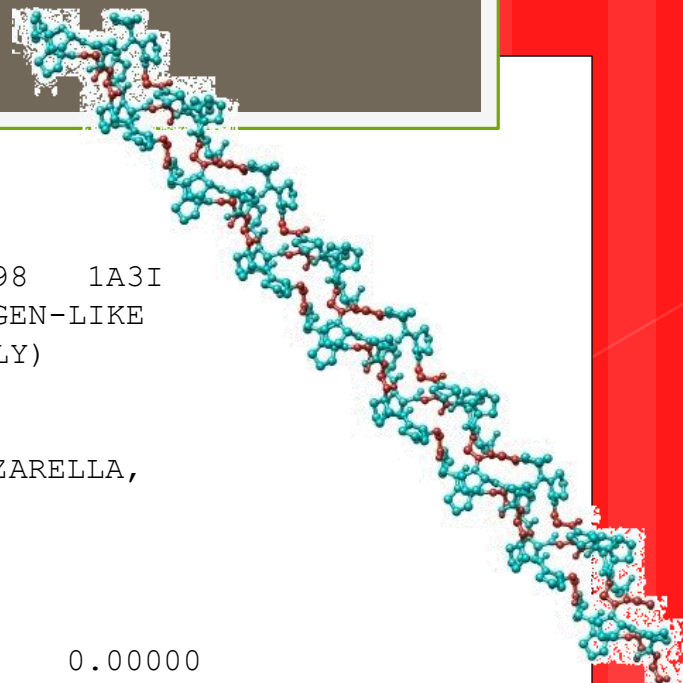
abx.2190	KEIAAALKQDNQVVKVEGITQDGENLTLSNEDFKLQKKIN
cas.2954	KEIAAALKQDNQVVKVEGITQDGENLTLSNEDFKLQKKIN
dex.897	KVIAKEANKATIOGNONIILLGTDSSGKKLKIDNEDVKIQ

```
abx.2190      VKSGDVTVKKTDTKTQTILDKIEPKSITIVKNEEETG*--
cas.2954      VKSGDVTVKKTDTKTQTILDKIEPKSITIVKNEEETG*--
dex.897       FKKLIOSKEIKVKHTDEETOEILNTIDTNKVMIIHENETE
```



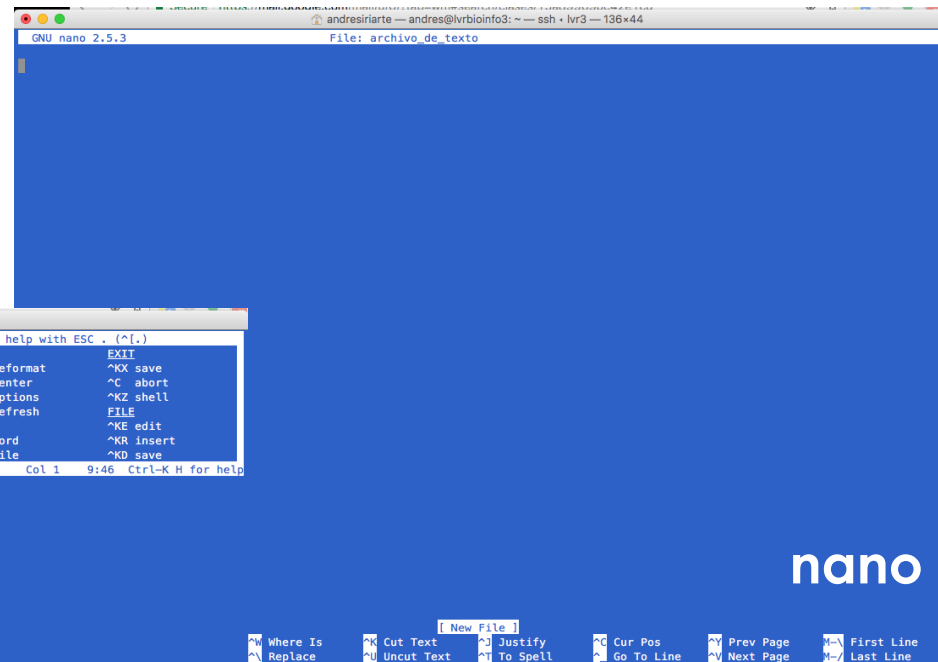
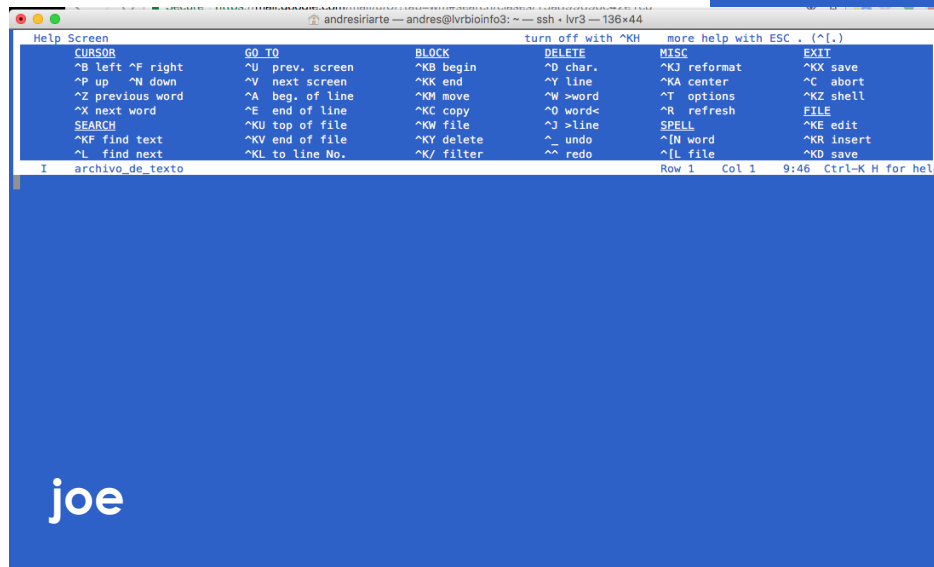
PDB:

HEADER EXTRACELLULAR MATRIX 22-JAN-98 1A3I
TITLE X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE 2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA X-RAY DIFFRACTION
AUTHOR R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR 2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350 BIOMT1 1 1.000000 0.000000 0.000000 0.00000
REMARK 350 BIOMT2 1 0.000000 1.000000 0.000000 0.00000
...
SEQRES 1 A 9 PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES 1 B 6 PRO PRO GLY PRO PRO GLY
SEQRES 1 C 6 PRO PRO GLY PRO PRO GLY
...
ATOM 1 N PRO A 1 8.316 21.206 21.530 1.00 17.44 N
ATOM 2 CA PRO A 1 7.608 20.729 20.336 1.00 17.44 C
ATOM 3 C PRO A 1 8.487 20.707 19.092 1.00 17.44 C
ATOM 4 O PRO A 1 9.466 21.457 19.005 1.00 17.44 O
ATOM 5 CB PRO A 1 6.460 21.723 20.211 1.00 22.26 C
...
HETATM 130 C ACY 401 3.682 22.541 11.236 1.00 21.19 C
HETATM 131 O ACY 401 2.807 23.097 10.553 1.00 21.19 O
HETATM 132 OXT ACY 401 4.306 23.101 12.291 1.00 21.19 O



Editores de texto en la línea de comandos

- joe, nano, vi, notas



Cuales son los comandos principales para la edición de textos.

grep o egrep (busca patrón)

- -c -> contar
- -w -> específicamente la palabra
- -v -> patrón opuesto
- -A 4 o -B 4 -> imprime 4 columnas luego del hit

sort (ordena)

- -k 3 -> ordena por valores en columna 3
- -R -> ordena al azar
- -r -> ordena reverso del alfabético (por defecto)
- -n -> ordena numericamente

Cuales son los comandos principales para la edición de textos.

sed (editor de texto)

- ◉ el comando esencial es “s” sustituir
 - ◉ sed 's/lindo/feo/g' archivo > archivo_modificado
 - ◉ sed 's/lindo//g' archivo > archivo_modificado
- ◉ -i genera cambio en el texto sin modificar el nombre
 - ◉ sed -i 's/lindo/feo/g' archivo (el archivo es modificado)
- ◉ puede utilizar *wildcards*: [0-9], ^..*, ^\$, [A-Z]...
- ◉ puede utilizar variables
 - ◉ sed 's/'\$var1'/lindo/g' archivo > archivo_modificado

Cuales son los comandos principales para la edición de textos.

head / tail (encabezado o pie)

- head -n 100 -> imprime las primeras 100 filas
- tail -n 2 -> imprime las ultimas 2 filas

wc (conteo)

- wc archivo -> cuenta el numero de líneas, palabras y caracteres en el archivo
- wc -l archivo -> cuenta líneas

tr (traduce)

- traduce o elimina



Cuales son los comandos principales para la edición de textos.

uniq (reporta u omite líneas repetidas consecutivas)

- `uniq -d archivo` -> imprime líneas repetidas
- `uniq -u archivo` -> imprime líneas únicas
- `uniq -w 10 archivo` -> considera repeticiones solo en los primeros 10 caracteres.
- `uniq -c archivo` -> reporta el conteo de cada una de las líneas repetidas.

uniq debe ser precedido por sort ojo!

Cuales son los comandos principales para la edición de textos.

awk (lenguaje para editar texto)

- `awk '{print $1}' archivo` -> imprime columna 1
- `awk '$3 > 4' archivo` -> imprime todas las filas y columnas en las que el valor de la columna 3 sea mayor a 4
- `awk '{ sum += $2; n++ } END { if (n > 0) print sum / n; }'`
- `awk '{for(i=1;i<=NF;i++) {sum[i] += $i; sumsq[i] += ($i)^2}} END {for (i=1;i<=NF;i++) {print "%f %f \n", sum[i]/NR, sqrt((sumsq[i]-sum[i]^2/NR)/NR)}}'`

Otros elementos útiles para el manejo de texto

echo (imprime)

- echo "lindo" > archivo -> imprime lindo y lo guarda en el archivo
- echo "\$Variable"

>, >>, 2>, |

- > re-dirige la salida estándar
- >> re-dirige la salida estándar y agrega en archivo
- 2> re-dirige el error estándar
- | concatena salidas y entradas estándar

Otros elementos útiles para el manejo de texto

\n, \t, \$, ^, “ ”, *, [a-z]

- \n salto de línea
- \t tabulador
- \$ fin de línea
- ^ principio de línea
- “ ” espacio
- * todos los caracteres, repetidos n veces
- [a-z] todas las letras minúsculas una vez
- [0-9] todos los números una vez
- [ACGhstr] cualquiera de los caracteres dentro de los paréntesis rectos