Useful shell commands

head/tail, cut, sort, uniq

Virginie Orgogozo

March 2011

Grep

Prints out the lines containing the characters

Options

-c
Shows only a count of the results

-v
Shows only the lines that do not match the pattern. Inverted search.

-i
ignore case

-E
Use regular expressions. Terms should be in quotes, use [] to indicate a character range, use [[:space:]] for \s, [[:digit:]] for \d.

-n
Show line number of the matches

Agrep (Approximate grep)

searches for a nearly exact match.

Options

-d "\>"
uses > as a delimiter between records rather than end-of-line

-B -y
returns only the best match
$agrep -B -y -d "\>" CYG FPexcerpt.fta

-2
returns results with up to this many mismatches between query and record. Maximum allowed is 8.

-l
only lists filenames that contain a match

-i
case-insensitive search

Useful tips

How to write tab or enter characters in the shell?

Press Ctrl+V first and then the special character.
"Enter**"** is represented by **"**^M**"**

How to search for negative numbers with grep ?

$grep "\-122" ctd.txt

Cut
Head/tail
Grep
Sort
Uniq

Exercice

From structure_1sl8.pdb
Obtain the number of amino acids

```
HEADER    LUMINESCENT PROTEIN                   05-MAR-04   1SL8

TITLE     CALCIUM-LOADED APO-AEQUORIN FROM AEQUOREA
VICTORIA
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: AEQUORIN 1;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
(...)
ATOM      1  N   ASN A  11      1.700   5.666
ATOM      2  CA  ASN A  11      2.196   7.022
ATOM      3  C   ASN A  11      1.537   7.599
ATOM      4  O   ASN A  11      1.077   8.750
ATOM      5  CB  ASN A  11      2.078   8.057
ATOM      6  CG  ASN A  11      2.982   7.755
(...)
```

```
ASN  11
ASN  11
ASN  11
ASN  11
ASN  11
ASN  11
ASN  11
ASN  11
PRO  12
PRO  12
PRO  12
PRO  12
PRO  12
PRO  12
PRO  12
PRO  12
LYS  13
LYS  13
LYS  13
LYS  13
LYS  13
(...)
```

```
ALA 113
ALA 125
ALA 133
ALA 138
ALA 181
ALA 189
ALA  42
ALA  52
ALA  57
ALA  63
ALA  66
ALA  71
ALA  92
ARG 108
ARG 152
ARG 169
ARG  17
ARG  32
ARG  59
ARG  90
ARG  98
ASN 102
ASN  11
ASN 123
(...)
```

```
 16 GLU
 15 GLY
 15 ASP
 13 LYS
 13 ILE
 13 ALA
 12 LEU
  9 SER
  8 VAL
  8 ASN
  8 ARG
  7 TYR
  7 THR
  7 PHE
  6 TRP
  6 GLN
  5 PRO
  5 MET
  5 HIS
  3 CYS
```

Exercice

From structure_1sl8.pdb
Obtain the number of amino acids

```
HEADER    LUMINESCENT PROTEIN                   05-MAR-04   1SL8

TITLE     CALCIUM-LOADED APO-AEQUORIN FROM AEQUOREA
VICTORIA
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: AEQUORIN 1;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
(...)
ATOM      1  N   ASN A  11       1.700   5.666
ATOM      2  CA  ASN A  11       2.196   7.022
ATOM      3  C   ASN A  11       1.537   7.599
ATOM      4  O   ASN A  11       1.077   8.750
ATOM      5  CB  ASN A  11       2.078   8.057
ATOM      6  CG  ASN A  11       2.982   7.755
(...)
```

```
ASN  11
ASN  11
ASN  11
ASN  11
ASN  11
ASN  11
ASN  11
ASN  11
PRO  12
PRO  12
PRO  12
PRO  12
PRO  12
PRO  12
PRO  12
LYS  13
LYS  13
LYS  13
LYS  13
LYS  13
(...)
```

```
ALA 113
ALA 125
ALA 133
ALA 138
ALA 181
ALA 189
ALA  42
ALA  52
ALA  57
ALA  63
ALA  66
ALA  71
ALA  92
ARG 108
ARG 152
ARG 169
ARG  17
ARG  32
ARG  59
ARG  90
ARG  98
(...)
```

```
 16 GLU
 15 GLY
 15 ASP
 13 LYS
 13 ILE
 13 ALA
 12 LEU
  9 SER
  8 VAL
  8 ASN
  8 ARG
  7 TYR
  7 THR
  7 PHE
  6 TRP
  6 GLN
  5 PRO
  5 MET
  5 HIS
  3 CYS
```

```
grep ATOM structure_1sl8.pdb |grep -v REMARK|cut -c 18-21,24-26|
uniq|sort|cut -c 1-3|uniq -c|sort -nr
```