Refresh

Protein function

BRENDA

Catalytic Site Atlas

Membrane proteins

Protein structure

awk programming

awk is a pattern-matching program for processing files, especially when they are databases (when each line has a simple field oriented layout).

The awk utility is a data extraction and reporting tool that uses a data-driven scripting language consisting of a set of actions to be taken against textual data for the purpose of producing formatted reports.

It is a very powerful program for handling large amount of data (especially parsing data files in bioinformatics).

awk is one of the earliest tools to appear in Unix and gained popularity as a way to add computational features to a Unix pipeline.

awk was created at Bell Labs in the 1970s, and its name is derived from the family names of its authors — Alfred Aho, Peter Weinberger, and Brian Kernighan.

What can we do with awk?

- Think of a text file as made up of records and fields in a textual database.
- Perform arithmetic and string operations.
- Use programming constructs such as loops and conditionals.
- Produce formatted reports.

AWK - the original from AT&T
NAWK - A newer, improved version from AT&T
GAWK - The Free Software foundation's version

- With nawk, you can also:
- Execute Unix commands from a script.
- Process the results of Unix commands.
- Process command-line arguments more gracefully.
- Work more easily with multiple input streams.
- Perform more powerful string substitutions (gawk)

awk one liners

Syntax

awk [options] ‘script’ var=value file(s)

Pattern-action statement

awk ‘pattern {*action*}’ file name

E.g. awk ‘NF>1 {*print $1*}’ abc.dat

awk [options] –f scriptfile var=value file(s)

-F *fs:* Use *fs* for the input field separator (the value of the FS predefined variable).

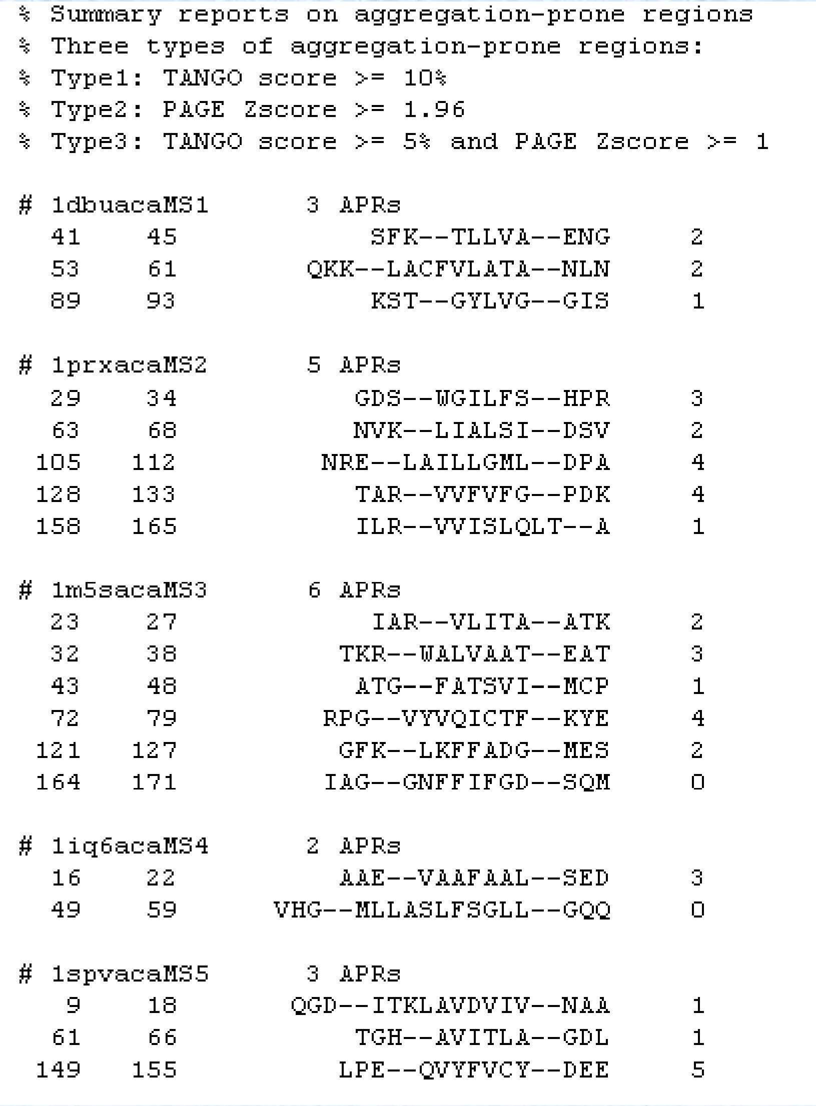
-f *program-file:* Read the awk program source from the file *program-file*, instead of from the first command line argument.

-v *var=val:* Assign the variable *var* the value *val* before program execution begins.

--: Signal the end of options.

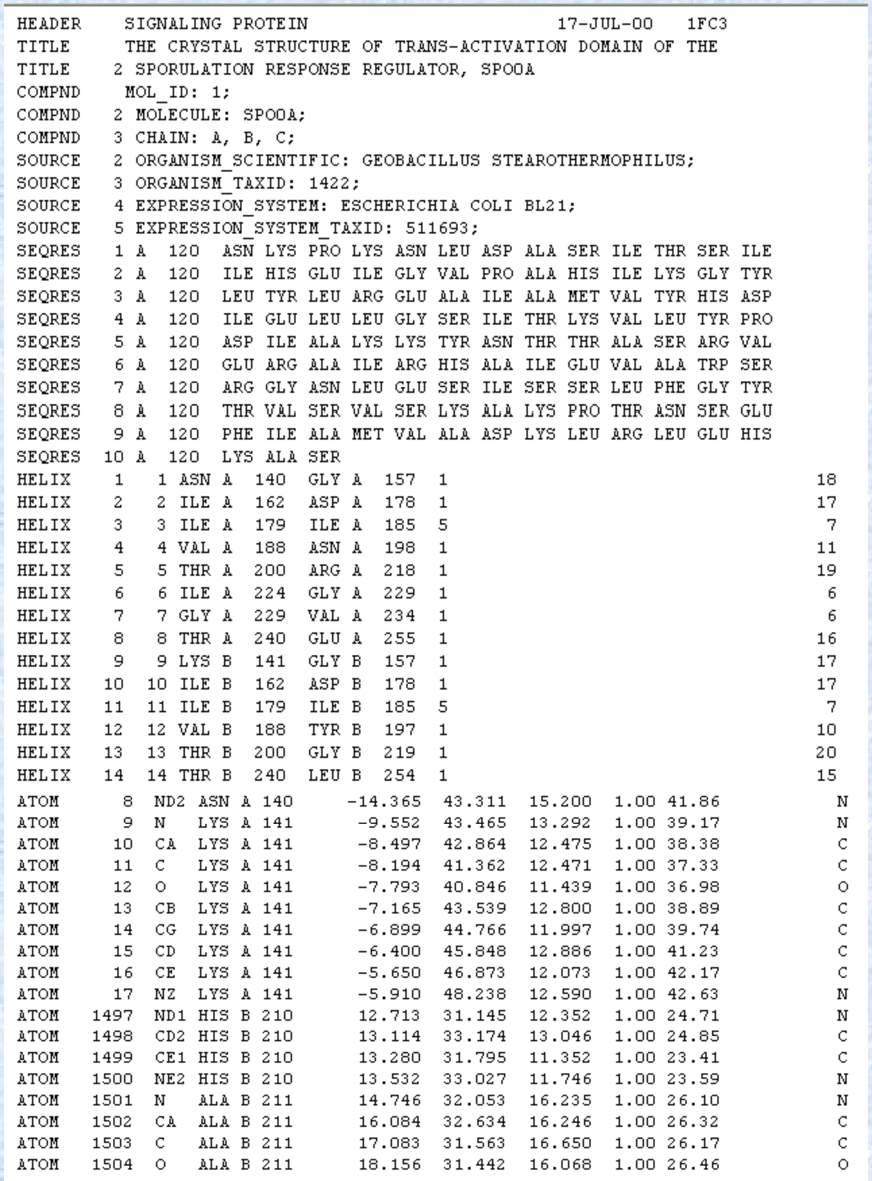
Sample input files

test1.dat



```
% Summary reports on aggregation-prone regions
% Three types of aggregation-prone regions:
% Type1: TANGO score >= 10%
% Type2: PAGE Zscore >= 1.96
% Type3: TANGO score >= 5% and PAGE Zscore >= 1

# 1dbuacaMS1       3 APRs
  41    45           SFK--TLLVA--ENG     2
  53    61        QKK--LACFVLATA--NLN    2
  89    93           KST--GYLVG--GIS     1

# 1prxacaMS2       5 APRs
  29    34          GDS--WGILFS--HPR     3
  63    68          NVK--LIALSI--DSV     2
 105   112        NRE--LAILLGML--DPA     4
 128   133          TAR--VVFVFG--PDK     4
 158   165         ILR--VVISLQLT--A      1

# 1m5sacaMS3       6 APRs
  23    27           IAR--VLITA--ATK     2
  32    38         TKR--WALVAAT--EAT     3
  43    48          ATG--FATSVI--MCP     1
  72    79        RPG--VYVQICTF--KYE     4
 121   127         GFK--LKFFADG--MES     2
 164   171        IAG--GNFFIFGD--SQM     0

# 1iq6acaMS4       2 APRs
  16    22         AAE--VAAFAAL--SED     3
  49    59     VHG--MLLASLFSGLL--GQQ     0

# 1spvacaMS5       3 APRs
   9    18      QGD--ITKLAVDVIV--NAA     1
  61    66          TGH--AVITLA--GDL     1
 149   155         LPE--QVYFVCY--DEE     5
```

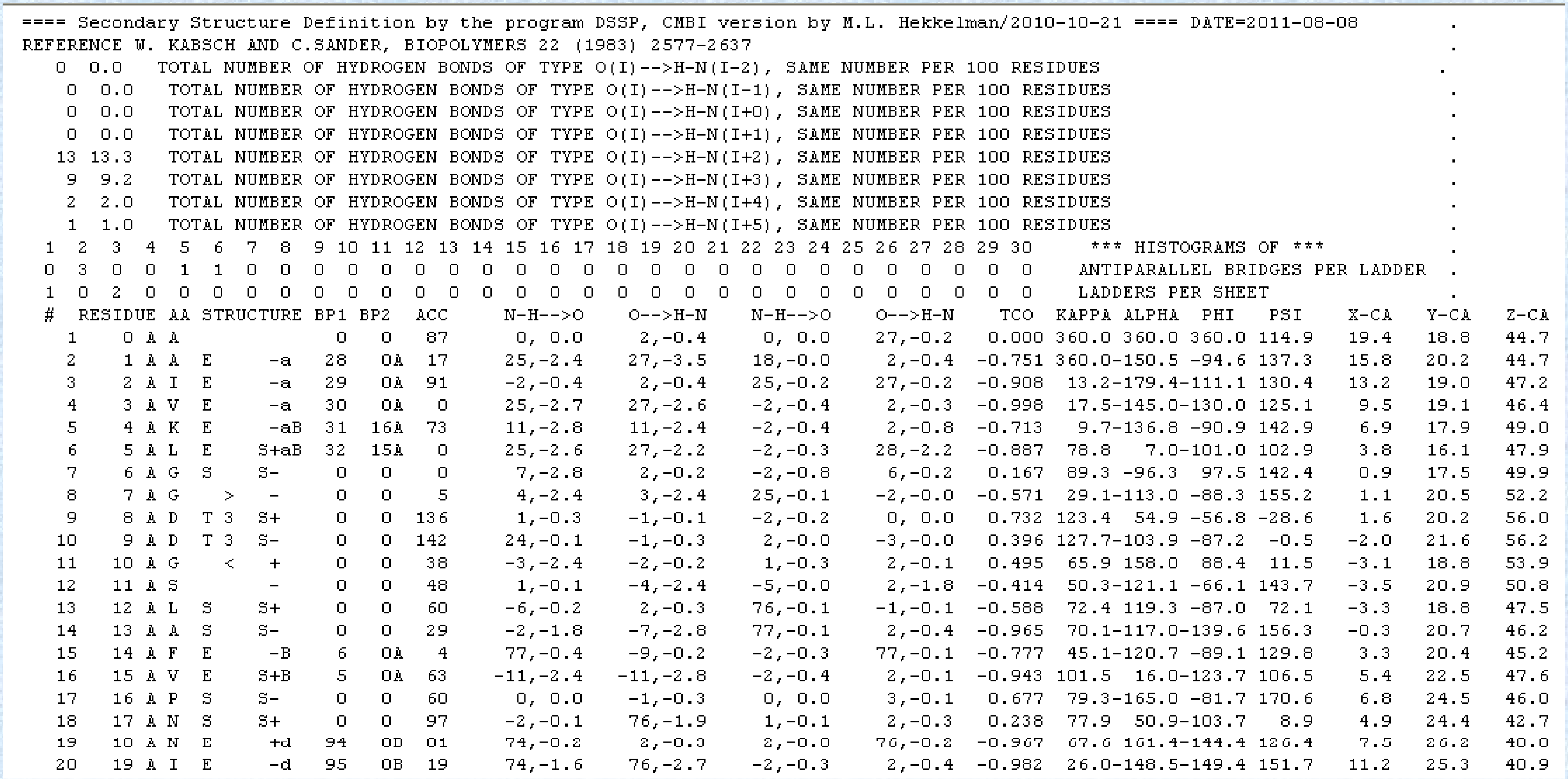
test2.dat



```
HEADER    SIGNALING PROTEIN                       17-JUL-00   1FC3
TITLE     THE CRYSTAL STRUCTURE OF TRANS-ACTIVATION DOMAIN OF THE
TITLE    2 SPORULATION RESPONSE REGULATOR, SPOOA
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: SPOOA;
COMPND   3 CHAIN: A, B, C;
SOURCE   2 ORGANISM_SCIENTIFIC: GEOBACILLUS STEAROTHERMOPHILUS;
SOURCE   3 ORGANISM_TAXID: 1422;
SOURCE   4 EXPRESSION_SYSTEM: ESCHERICHIA COLI BL21;
SOURCE   5 EXPRESSION_SYSTEM_TAXID: 511693;
SEQRES   1 A  120  ASN LYS PRO LYS ASN LEU ASP ALA SER ILE THR SER ILE
SEQRES   2 A  120  ILE HIS GLU ILE GLY VAL PRO ALA HIS ILE LYS GLY TYR
SEQRES   3 A  120  LEU TYR LEU ARG GLU ALA ILE ALA MET VAL TYR HIS ASP
SEQRES   4 A  120  ILE GLU LEU LEU GLY SER ILE THR LYS VAL LEU TYR PRO
SEQRES   5 A  120  ASP ILE ALA LYS LYS TYR ASN THR THR ALA SER ARG VAL
SEQRES   6 A  120  GLU ARG ALA ILE ARG HIS ALA ILE GLU VAL ALA TRP SER
SEQRES   7 A  120  ARG GLY ASN LEU GLU SER ILE SER SER LEU PHE GLY TYR
SEQRES   8 A  120  THR VAL SER VAL SER LYS ALA LYS PRO THR ASN SER GLU
SEQRES   9 A  120  PHE ILE ALA MET VAL ALA ASP LYS LEU ARG LEU GLU HIS
SEQRES  10 A  120  LYS ALA SER
HELIX    1   1 ASN A  140  GLY A  157  1                                  18
HELIX    2   2 ILE A  162  ASP A  178  1                                  17
HELIX    3   3 ILE A  179  ILE A  185  5                                   7
HELIX    4   4 VAL A  188  ASN A  198  1                                  11
HELIX    5   5 THR A  200  ARG A  218  1                                  19
HELIX    6   6 ILE A  224  GLY A  229  1                                   6
HELIX    7   7 GLY A  229  VAL A  234  1                                   6
HELIX    8   8 THR A  240  GLU A  255  1                                  16
HELIX    9   9 LYS B  141  GLY B  157  1                                  17
HELIX   10  10 ILE B  162  ASP B  178  1                                  17
HELIX   11  11 ILE B  179  ILE B  185  5                                   7
HELIX   12  12 VAL B  188  TYR B  197  1                                  10
HELIX   13  13 THR B  200  GLY B  219  1                                  20
HELIX   14  14 THR B  240  LEU B  254  1                                  15
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86           N
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

Sample input files

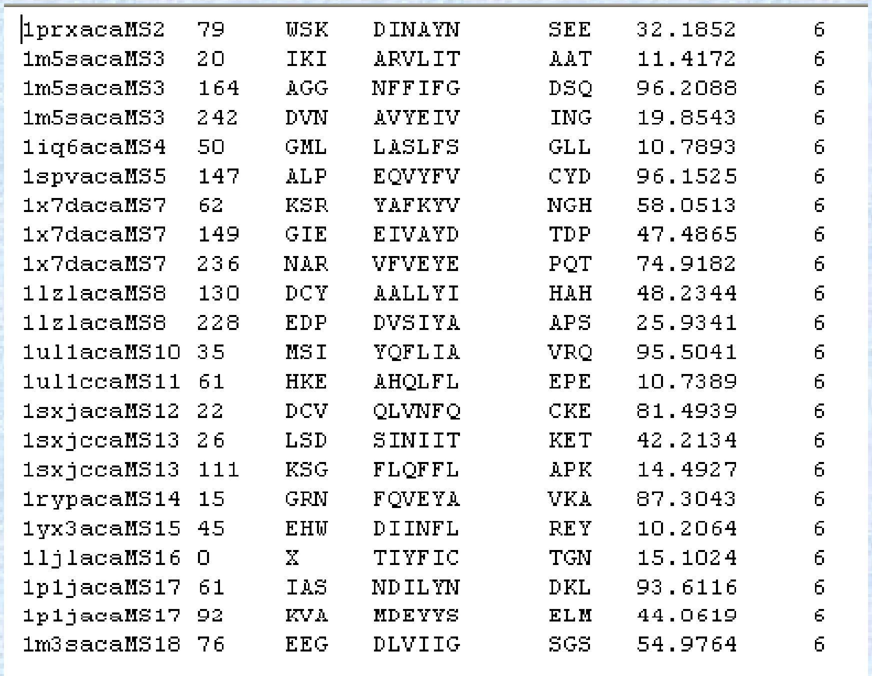
test3.dat



```
==== Secondary Structure Definition by the program DSSP, CMBI version by M.L. Hekkelman/2010-10-21 ==== DATE=2011-08-08        .
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637                                                              .
  0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES                               .
   0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES                              .
   0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES                              .
   0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES                              .
  13 13.3   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES                              .
   9  9.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES                              .
   2  2.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES                              .
   1  1.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES                              .
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30     *** HISTOGRAMS OF ***           .
  0  3  0  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    ANTIPARALLEL BRIDGES PER LADDER  .
  1  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    LADDERS PER SHEET                .
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
    1    0 A A              0   0   87      0, 0.0     2,-0.4     0, 0.0    27,-0.2   0.000 360.0 360.0 360.0 114.9   19.4   18.8   44.7
    2    1 A A  E     -a   28  0A   17     25,-2.4    27,-3.5    18,-0.0     2,-0.4  -0.751 360.0-150.5 -94.6 137.3   15.8   20.2   44.7
    3    2 A I  E     -a   29  0A   91     -2,-0.4     2,-0.4    25,-0.2    27,-0.2  -0.908  13.2-179.4-111.1 130.4   13.2   19.0   47.2
    4    3 A V  E     -a   30  0A    0     25,-2.7    27,-2.6    -2,-0.4     2,-0.3  -0.998  17.5-145.0-130.0 125.1    9.5   19.1   46.4
    5    4 A K  E     -aB  31 16A   73     11,-2.8    11,-2.4    -2,-0.4     2,-0.8  -0.713   9.7-136.8 -90.9 142.9    6.9   17.9   49.0
    6    5 A L  E    S+aB  32 15A    0     25,-2.6    27,-2.2    -2,-0.3    28,-2.2  -0.887  78.8   7.0-101.0 102.9    3.8   16.1   47.9
    7    6 A G  S    S-     0   0    0      7,-2.8     2,-0.2    -2,-0.8     6,-0.2   0.167  89.3 -96.3  97.5 142.4    0.9   17.5   49.9
    8    7 A G    >  -      0   0    5      4,-2.4     3,-2.4    25,-0.1    -2,-0.0  -0.571  29.1-113.0 -88.3 155.2    1.1   20.5   52.2
    9    8 A D  T 3  S+     0   0  136      1,-0.3    -1,-0.1    -2,-0.2     0, 0.0   0.732 123.4  54.9 -56.8 -28.6    1.6   20.2   56.0
   10    9 A D  T 3  S-     0   0  142     24,-0.1    -1,-0.3     2,-0.0    -3,-0.0   0.396 127.7-103.9 -87.2  -0.5   -2.0   21.6   56.2
   11   10 A G    <  +      0   0   38     -3,-2.4    -2,-0.2     1,-0.3     2,-0.1   0.495  65.9 158.0  88.4  11.5   -3.1   18.8   53.9
   12   11 A S       -      0   0   48      1,-0.1    -4,-2.4    -5,-0.0     2,-1.8  -0.414  50.3-121.1 -66.1 143.7   -3.5   20.9   50.8
   13   12 A L  S    S+     0   0   60     -6,-0.2     2,-0.3    76,-0.1    -1,-0.1  -0.588  72.4 119.3 -87.0  72.1   -3.3   18.8   47.5
   14   13 A A  S    S-     0   0   29     -2,-1.8    -7,-2.8    77,-0.1     2,-0.4  -0.965  70.1-117.0-139.6 156.3   -0.3   20.7   46.2
   15   14 A F  E     -B    6  0A    4     77,-0.4    -9,-0.2    -2,-0.3    77,-0.1  -0.777  45.1-120.7 -89.1 129.8    3.3   20.4   45.2
   16   15 A V  E    S+B    5  0A   63    -11,-2.4   -11,-2.8    -2,-0.4     2,-0.1  -0.943 101.5  16.0-123.7 106.5    5.4   22.5   47.6
   17   16 A P  S    S-     0   0   60      0, 0.0    -1,-0.3     0, 0.0     3,-0.1   0.677  79.3-165.0 -81.7 170.6    6.8   24.5   46.0
   18   17 A N  S    S+     0   0   97     -2,-0.1    76,-1.9     1,-0.1     2,-0.3   0.238  77.9  50.9-103.7   8.9    4.9   24.4   42.7
   19   18 A N  E     +d   94  0D   81     74,-0.2     2,-0.3     2,-0.0    76,-0.2  -0.967  67.6 161.4-144.4 126.4    7.5   26.2   40.0
   20   19 A I  E     -d   95  0B   19     74,-1.6    76,-2.7    -2,-0.3     2,-0.4  -0.982  26.0-148.5-149.4 151.7   11.2   25.3   40.9
```

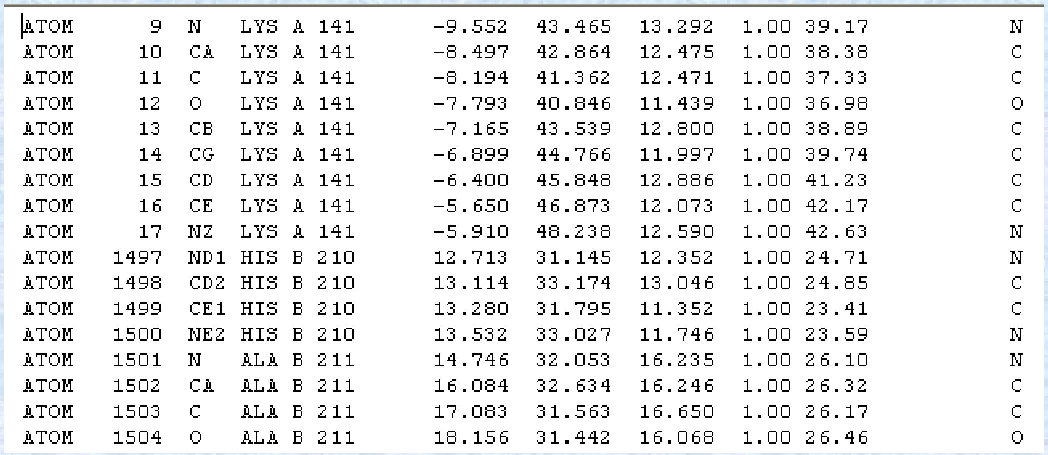
Sample input files

test4.dat



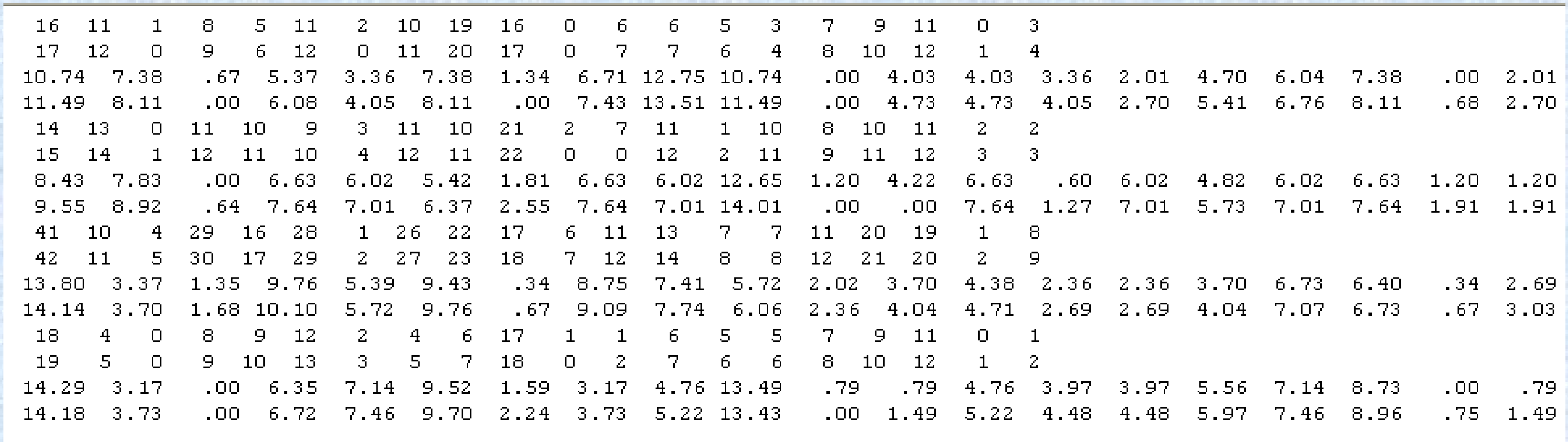
```
1prxacaMS2  79    WSK   DINAYN      SEE   32.1852      6
1m5sacaMS3  20    IKI   ARVLIT      AAT   11.4172      6
1m5sacaMS3  164   AGG   NFFIFG      DSQ   96.2088      6
1m5sacaMS3  242   DVN   AVYEIV      ING   19.8543      6
1iq6acaMS4  50    GML   LASLFS      GLL   10.7893      6
1spvacaMS5  147   ALP   EQVYFV      CYD   96.1525      6
1x7dacaMS7  62    KSR   YAFKYV      NGH   58.0513      6
1x7dacaMS7  149   GIE   EIVAYD      TDP   47.4865      6
1x7dacaMS7  236   NAR   VFVEYE      PQT   74.9182      6
1lzlacaMS8  130   DCY   AALLYI      HAH   48.2344      6
1lzlacaMS8  228   EDP   DVSIYA      APS   25.9341      6
1ul1acaMS10 35    MSI   YQFLIA      VRQ   95.5041      6
1ul1ccaMS11 61    HKE   AHQLFL      EPE   10.7389      6
1sxjacaMS12 22    DCV   QLVNFQ      CKE   81.4939      6
1sxjccaMS13 26    LSD   SINIIT      KET   42.2134      6
1sxjccaMS13 111   KSG   FLQFFL      APK   14.4927      6
1rypacaMS14 15    GRN   FQVEYA      VKA   87.3043      6
1yx3acaMS15 45    EHW   DIINFL      REY   10.2064      6
1ljlacaMS16 0     X     TIYFIC      TGN   15.1024      6
1p1jacaMS17 61    IAS   NDILYN      DKL   93.6116      6
1p1jacaMS17 92    KVA   MDEYYS      ELM   44.0619      6
1m3sacaMS18 76    EEG   DLVIIG      SGS   54.9764      6
```

test6.dat



```
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

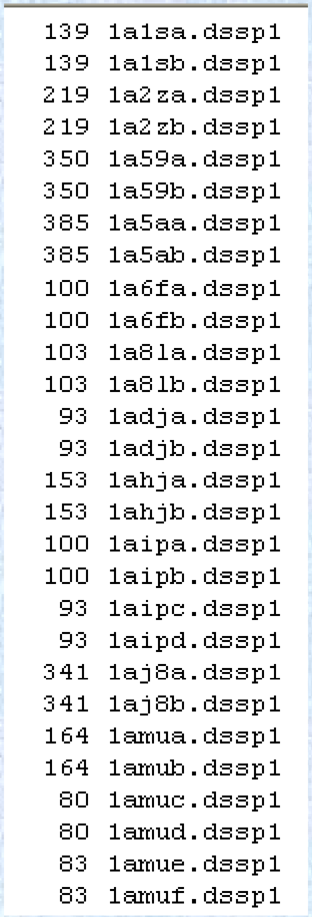
test5.dat



```
    16    11     1     8     5    11     2    10    19    16     0     6     6     5     3     7     9    11     0     3
    17    12     0     9     6    12     0    11    20    17     0     7     7     6     4     8    10    12     1     4
 10.74  7.38   .67  5.37  3.36  7.38  1.34  6.71 12.75 10.74   .00  4.03  4.03  3.36  2.01  4.70  6.04  7.38   .00  2.01
 11.49  8.11   .00  6.08  4.05  8.11   .00  7.43 13.51 11.49   .00  4.73  4.73  4.05  2.70  5.41  6.76  8.11   .68  2.70
    14    13     0    11    10     9     3    11    10    21     2     7    11     1    10     8    10    11     2     2
    15    14     1    12    11    10     4    12    11    22     0     0    12     2    11     9    11    12     3     3
  8.43  7.83   .00  6.63  6.02  5.42  1.81  6.63  6.02 12.65  1.20  4.22  6.63   .60  6.02  4.82  6.02  6.63  1.20  1.20
  9.55  8.92   .64  7.64  7.01  6.37  2.55  7.64  7.01 14.01   .00   .00  7.64  1.27  7.01  5.73  7.01  7.64  1.91  1.91
    41    10     4    29    16    28     1    26    22    17     6    11    13     7     7    11    20    19     1     8
    42    11     5    30    17    29     2    27    23    18     7    12    14     8     8    12    21    20     2     9
 13.80  3.37  1.35  9.76  5.39  9.43   .34  8.75  7.41  5.72  2.02  3.70  4.38  2.36  2.36  3.70  6.73  6.40   .34  2.69
 14.14  3.70  1.68 10.10  5.72  9.76   .67  9.09  7.74  6.06  2.36  4.04  4.71  2.69  2.69  4.04  7.07  6.73   .67  3.03
    18     4     0     8     9    12     2     4     6    17     1     1     6     5     5     7     9    11     0     1
    19     5     0     9    10    13     3     5     7    18     0     2     7     6     6     8    10    12     1     2
 14.29  3.17   .00  6.35  7.14  9.52  1.59  3.17  4.76 13.49   .79   .79  4.76  3.97  3.97  5.56  7.14  8.73   .00   .79
 14.18  3.73   .00  6.72  7.46  9.70  2.24  3.73  5.22 13.43   .00  1.49  5.22  4.48  4.48  5.97  7.46  8.96   .75  1.49
```

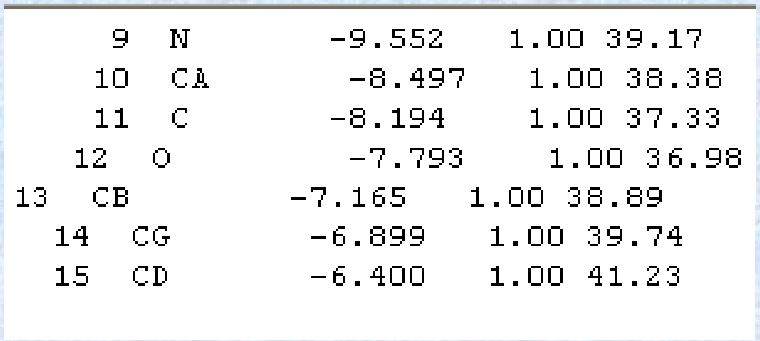
Sample input files

test8.dat



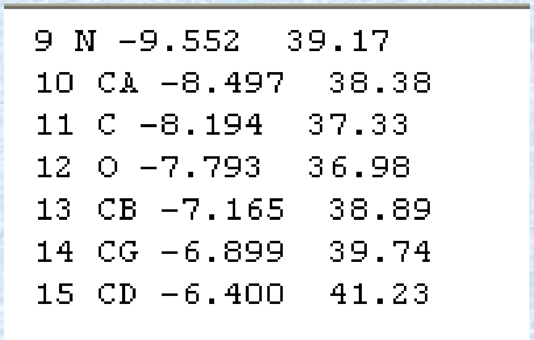
```
139 1a1sa.dssp1
139 1a1sb.dssp1
219 1a2za.dssp1
219 1a2zb.dssp1
350 1a59a.dssp1
350 1a59b.dssp1
385 1a5aa.dssp1
385 1a5ab.dssp1
100 1a6fa.dssp1
100 1a6fb.dssp1
103 1a8la.dssp1
103 1a8lb.dssp1
 93 1adja.dssp1
 93 1adjb.dssp1
153 1ahja.dssp1
153 1ahjb.dssp1
100 1aipa.dssp1
100 1aipb.dssp1
 93 1aipc.dssp1
 93 1aipd.dssp1
341 1aj8a.dssp1
341 1aj8b.dssp1
164 1amua.dssp1
164 1amub.dssp1
 80 1amuc.dssp1
 80 1amud.dssp1
 83 1amue.dssp1
 83 1amuf.dssp1
```

test9.dat



```
     9  N        -9.552    1.00 39.17
    10  CA        -8.497    1.00 38.38
    11  C        -8.194     1.00 37.33
   12  O          -7.793    1.00 36.98
13  CB         -7.165    1.00 38.89
  14  CG         -6.899    1.00 39.74
  14  CD         -6.400    1.00 41.23
```

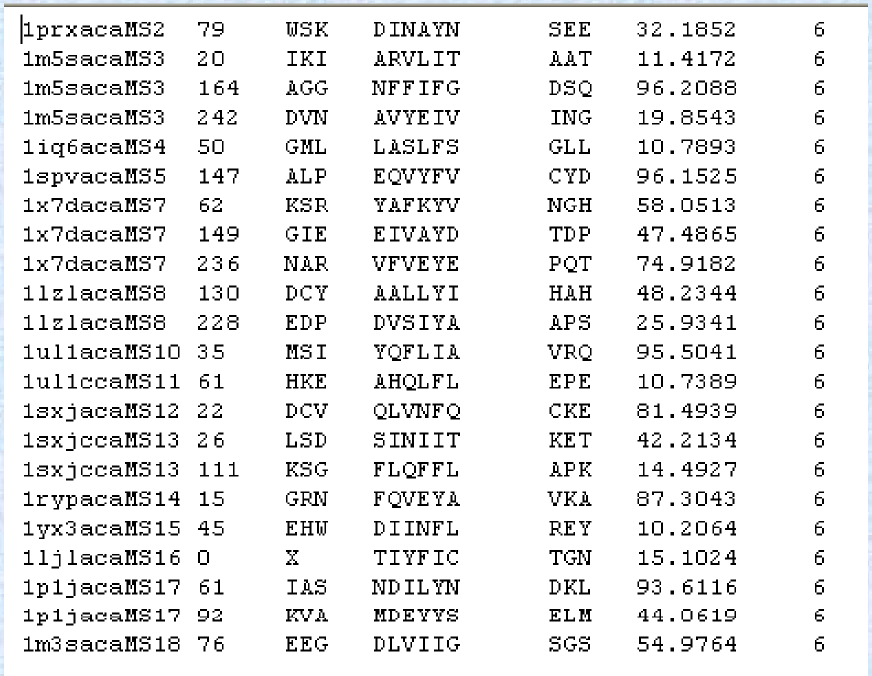
test10.dat



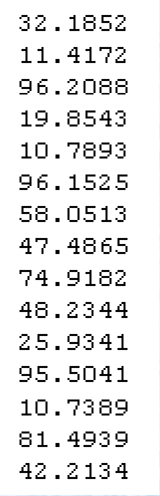
```
9  N -9.552  39.17
10 CA -8.497  38.38
11 C -8.194  37.33
12 O -7.793  36.98
13 CB -7.165  38.89
14 CG -6.899  39.74
15 CD -6.400  41.23
```

1. Print column 6

awk '{print $6}' test4.dat



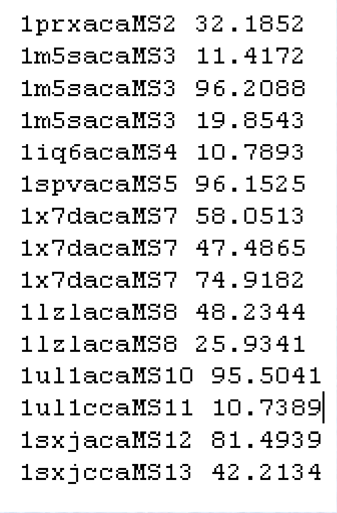
```
1prxacaMS2   79    WSK    DINAYN      SEE    32.1852      6
1m5sacaMS3   20    IKI    ARVLIT      AAT    11.4172      6
1m5sacaMS3   164   AGG    NFFIFG      DSQ    96.2088      6
1m5sacaMS3   242   DVN    AVYEIV      ING    19.8543      6
1iq6acaMS4   50    GML    LASLFS      GLL    10.7893      6
1spvacaMS5   147   ALP    EQVYFV      CYD    96.1525      6
1x7dacaMS7   62    KSR    YAFKYV      NGH    58.0513      6
1x7dacaMS7   149   GIE    EIVAYD      TDP    47.4865      6
1x7dacaMS7   236   NAR    VFVEYE      PQT    74.9182      6
1lzlacaMS8   130   DCY    AALLYI      HAH    48.2344      6
1lzlacaMS8   228   EDP    DVSIYA      APS    25.9341      6
1ul1acaMS10  35    MSI    YQFLIA      VRQ    95.5041      6
1ul1ccaMS11  61    HKE    AHQLFL      EPE    10.7389      6
1sxjacaMS12  22    DCV    QLVNFQ      CKE    81.4939      6
1sxjccaMS13  26    LSD    SINIIT      KET    42.2134      6
1sxjccaMS13  111   KSG    FLQFFL      APK    14.4927      6
1rypacaMS14  15    GRN    FQVEYA      VKA    87.3043      6
1yx3acaMS15  45    EHW    DIINFL      REY    10.2064      6
1ljlacaMS16  0     X      TIYFIC      TGN    15.1024      6
1p1jacaMS17  61    IAS    NDILYN      DKL    93.6116      6
1p1jacaMS17  92    KVA    MDEYYS      ELM    44.0619      6
1m3sacaMS18  76    EEG    DLVIIG      SGS    54.9764      6
```



```
32.1852
11.4172
96.2088
19.8543
10.7893
96.1525
58.0513
47.4865
74.9182
48.2344
25.9341
95.5041
10.7389
81.4939
42.2134
```

2. Print columns 1 and 6

awk '{print $1,$6}' test4.dat



```
1prxacaMS2 32.1852
1m5sacaMS3 11.4172
1m5sacaMS3 96.2088
1m5sacaMS3 19.8543
1iq6acaMS4 10.7893
1spvacaMS5 96.1525
1x7dacaMS7 58.0513
1x7dacaMS7 47.4865
1x7dacaMS7 74.9182
1lzlacaMS8 48.2344
1lzlacaMS8 25.9341
1ul1acaMS10 95.5041
1ul1ccaMS11 10.7389
1sxjacaMS12 81.4939
1sxjccaMS13 42.2134
```

3. Print in reverse order (columns 6 and 1)

awk '{print $6, $1}' test4.dat

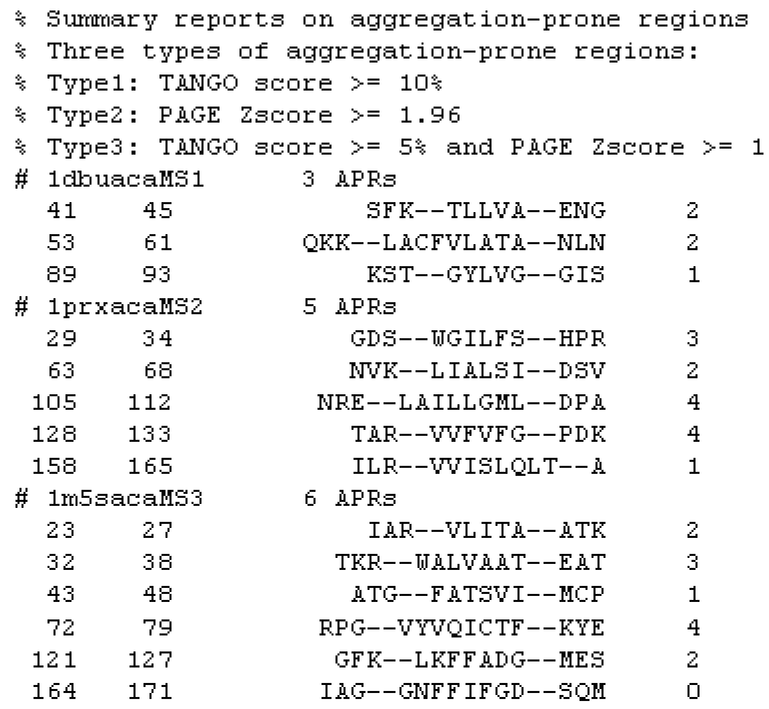
4. Write the results in a file

awk '{print $1" "$6}' test4.dat > test1.result

5. Delete empty lines

NF: number of fields

```
awk ' NF>1 {print}' test1.dat
```

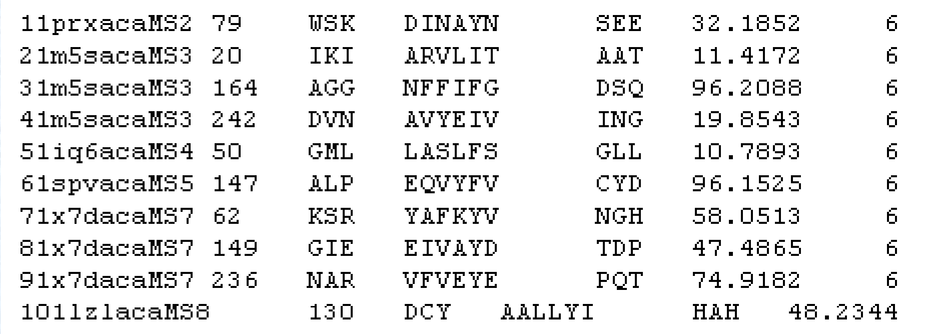


```
% Summary reports on aggregation-prone regions
% Three types of aggregation-prone regions:
% Type1: TANGO score >= 10%
% Type2: PAGE Zscore >= 1.96
% Type3: TANGO score >= 5% and PAGE Zscore >= 1
# 1dbuacaMS1      3 APRs
  41    45          SFK--TLLVA--ENG     2
  53    61      QKK--LACFVLATA--NLN     2
  89    93          KST--GYLVG--GIS     1
# 1prxacaMS2      5 APRs
  29    34         GDS--WGILFS--HPR     3
  63    68         NVK--LIALSI--DSV     2
 105   112       NRE--LAILLGML--DPA     4
 128   133         TAR--VVFVFG--PDK     4
 158   165       ILR--VVISLQLT--A       1
# 1m5sacaMS3      6 APRs
  23    27          IAR--VLITA--ATK     2
  32    38        TKR--WALVAAT--EAT     3
  43    48         ATG--FATSVI--MCP     1
  72    79       RPG--VYVQICTF--KYE     4
 121   127        GFK--LKFFADG--MES     2
 164   171       IAG--GNFFIFGD--SQM     0
```

6. Number each line

FNR: file line number

```
awk '{print FNR $0}' test4.dat
```

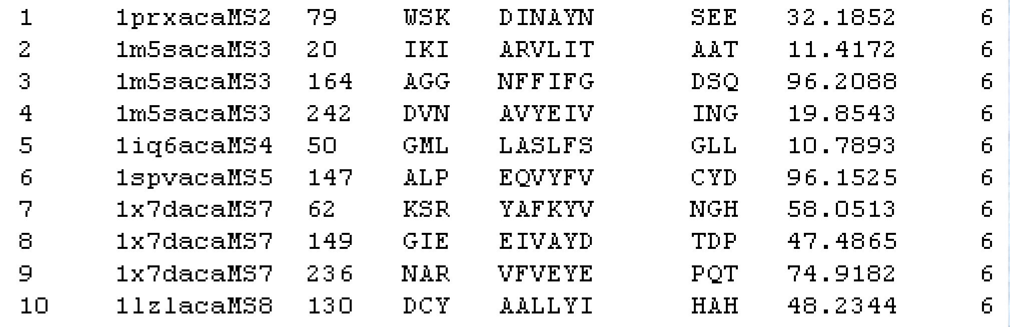


```
11prxacaMS2 79     WSK    DINAYN      SEE    32.1852      6
21m5sacaMS3 20     IKI    ARVLIT      AAT    11.4172      6
31m5sacaMS3 164    AGG    NFFIFG      DSQ    96.2088      6
41m5sacaMS3 242    DVN    AVYEIV      ING    19.8543      6
51iq6acaMS4 50     GML    LASLFS      GLL    10.7893      6
61spvacaMS5 147    ALP    EQVYFV      CYD    96.1525      6
71x7dacaMS7 62     KSR    YAFKYV      NGH    58.0513      6
81x7dacaMS7 149    GIE    EIVAYD      TDP    47.4865      6
91x7dacaMS7 236    NAR    VFVEYE      PQT    74.9182      6
1011zlacaMS8       130    DCY    AALLYI        HAH    48.2344
```

7. Number each line with tab

```
awk '{print FNR "\t" $0}' test4.dat
```

$0 denotes all fields



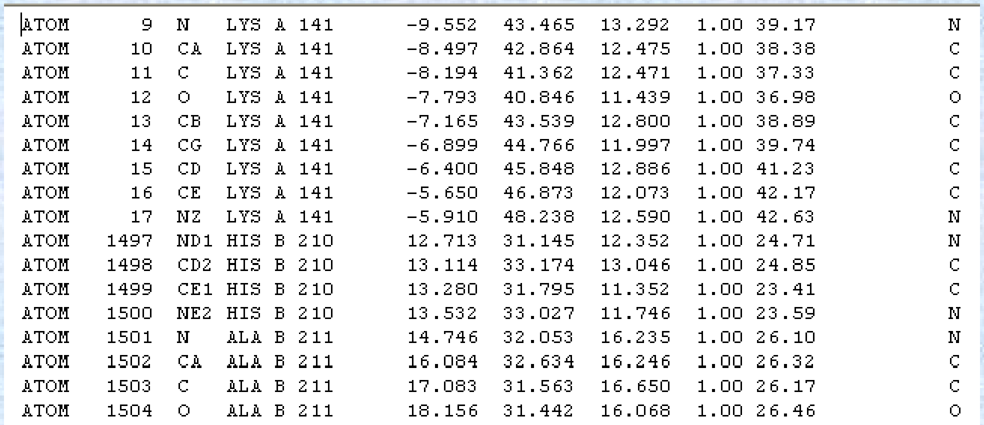
```
1      1prxacaMS2  79     WSK    DINAYN      SEE    32.1852      6
2      1m5sacaMS3  20     IKI    ARVLIT      AAT    11.4172      6
3      1m5sacaMS3  164    AGG    NFFIFG      DSQ    96.2088      6
4      1m5sacaMS3  242    DVN    AVYEIV      ING    19.8543      6
5      1iq6acaMS4  50     GML    LASLFS      GLL    10.7893      6
6      1spvacaMS5  147    ALP    EQVYFV      CYD    96.1525      6
7      1x7dacaMS7  62     KSR    YAFKYV      NGH    58.0513      6
8      1x7dacaMS7  149    GIE    EIVAYD      TDP    47.4865      6
9      1x7dacaMS7  236    NAR    VFVEYE      PQT    74.9182      6
10     1lzlacaMS8  130    DCY    AALLYI      HAH    48.2344      6
```

8. Count lines (similar to wc -l)

awk 'END {print NR}' test4.dat

NR: line number

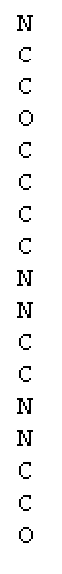
22



```
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

9. Print the last field of each line

awk '{print $NF}' test6.dat



```
N
C
C
O
C
C
C
C
N
N
C
C
N
N
C
C
O
```

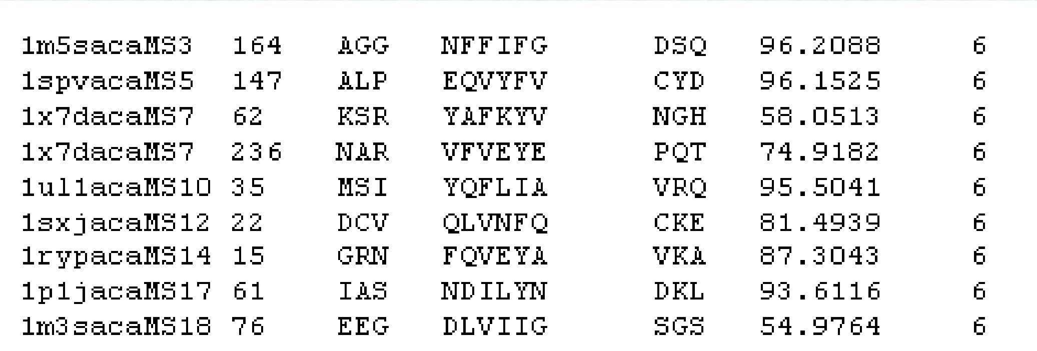
10. Print the last field of last line

awk ' END {print $NF}' test6.dat

O

11. Print every line, where the value of the 6th field is more than 50

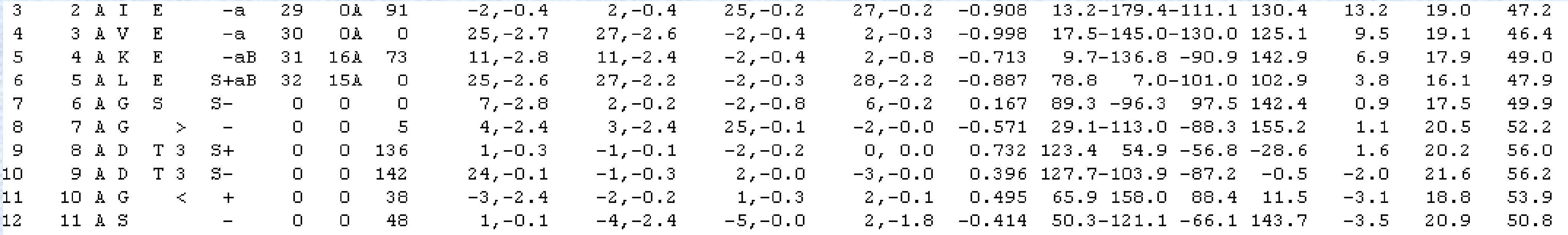
awk '$6 > 50 {print}' test4.dat



```
1m5sacaMS3   164   AGG   NFFIFG   DSQ   96.2088   6
1spvacaMS5   147   ALP   EQVYFV   CYD   96.1525   6
1x7dacaMS7   62    KSR   YAFKYV   NGH   58.0513   6
1x7dacaMS7   236   NAR   VFVEYE   PQT   74.9182   6
1ul1acaMS10  35    MSI   YQFLIA   VRQ   95.5041   6
1sxjacaMS12  22    DCV   QLVNFQ   CKE   81.4939   6
1rypacaMS14  15    GRN   FQVEYA   VKA   87.3043   6
1p1jacaMS17  61    IAS   NDILYN   DKL   93.6116   6
1m3sacaMS18  76    EEG   DLVIIG   SGS   54.9764   6
```

12. print the lines starting from 15

awk ' NR > 14 {print}' test3.dat



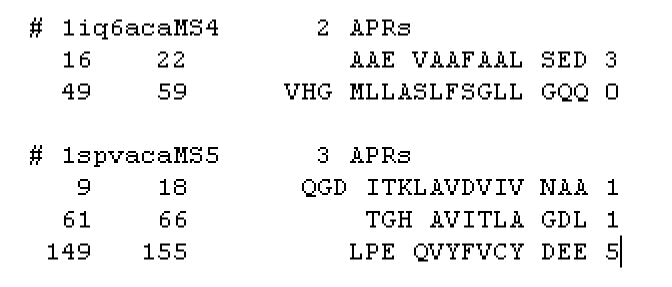
```
3     2 A I  E     -a   29   0A  91      -2,-0.4     2,-0.4    25,-0.2     27,-0.2  -0.908  13.2-179.4-111.1 130.4     13.2    19.0    47.2
4     3 A V  E     -a   30   0A   0      25,-2.7    27,-2.6    -2,-0.4      2,-0.3  -0.998  17.5-145.0-130.0 125.1      9.5    19.1    46.4
5     4 A K  E     -aB  31  16A  73      11,-2.8    11,-2.4    -2,-0.4      2,-0.8  -0.713   9.7-136.8 -90.9 142.9      6.9    17.9    49.0
6     5 A L  E    S+aB  32  15A   0      25,-2.6    27,-2.2    -2,-0.3     28,-2.2  -0.887  78.8   7.0-101.0 102.9      3.8    16.1    47.9
7     6 A G  S    S-     0   0    0       7,-2.8     2,-0.2    -2,-0.8      6,-0.2   0.167  89.3 -96.3  97.5 142.4      0.9    17.5    49.9
8     7 A G     >  -     0   0    5       4,-2.4     3,-2.4    25,-0.1     -2,-0.0  -0.571  29.1-113.0 -88.3 155.2      1.1    20.5    52.2
9     8 A D  T 3  S+     0   0  136       1,-0.3    -1,-0.1    -2,-0.2      0, 0.0   0.732 123.4  54.9 -56.8 -28.6      1.6    20.2    56.0
10    9 A D  T 3  S-     0   0  142      24,-0.1    -1,-0.3     2,-0.0     -3,-0.0   0.396 127.7-103.9 -87.2  -0.5     -2.0    21.6    56.2
11   10 A G     <  +     0   0   38      -3,-2.4    -2,-0.2     1,-0.3      2,-0.1   0.495  65.9 158.0  88.4  11.5     -3.1    18.8    53.9
12   11 A S        -     0   0   48       1,-0.1    -4,-2.4    -5,-0.0      2,-1.8  -0.414  50.3-121.1 -66.1 143.7     -3.5    20.9    50.8
```

13. Separate with field separator

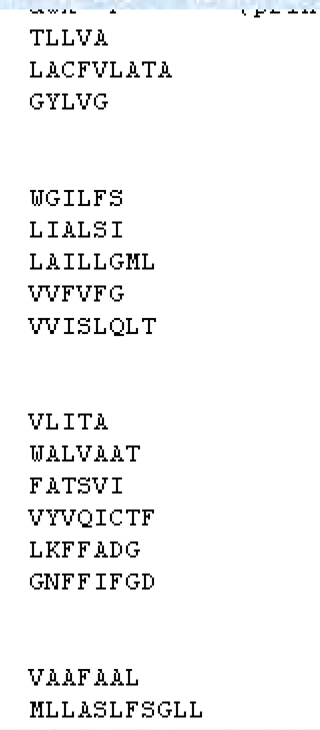
awk -F "--" '{print $2}' test1.dat

14. Separate with field separator

awk -F "--" '{print $1" "$2" "$3}' test1.dat



```
# 1iq6acaMS4      2 APRs
  16     22          AAE VAAFAAL SED 3
  49     59      VHG MLLASLFSGLL GQQ 0

# 1spvacaMS5      3 APRs
   9     18        QGD ITKLAVDVIV NAA 1
  61     66            TGH AVITLA GDL 1
 149    155            LPE QVYFVCY DEE 5
```

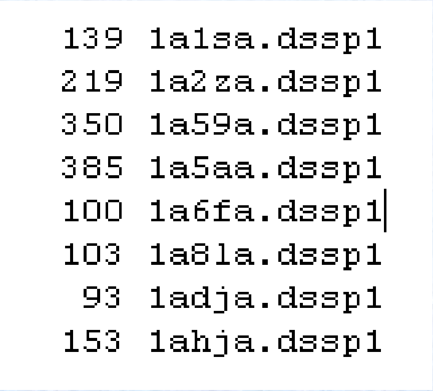


```
TLLVA
LACFVLATA
GYLVG


WGILFS
LIALSI
LAILLGML
VVFVFG
VVISLQLT


VLITA
WALVAAT
FATSVI
VYVQICTF
LKFFADG
GNFFIFGD


VAAFAAL
MLLASLFSGLL
```

M. Michael Gromiha, IIT Madras,

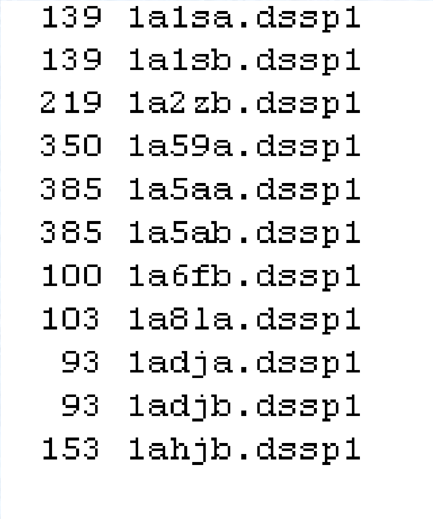
15. Print alternate lines

awk 'NR%2' test8.dat

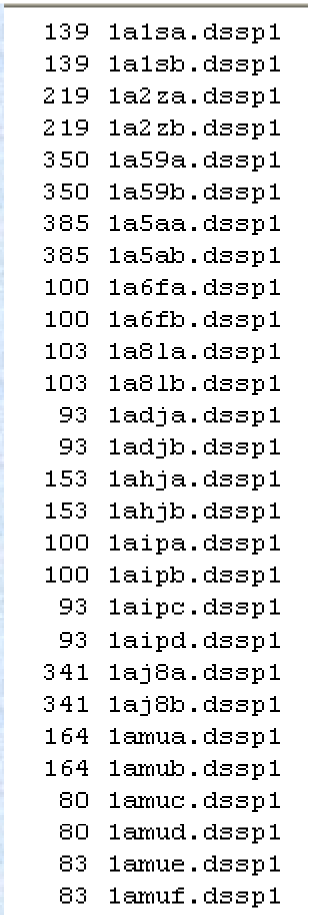


```
139 1a1sa.dssp1
219 1a2za.dssp1
350 1a59a.dssp1
385 1a5aa.dssp1
100 1a6fa.dssp1
103 1a8la.dssp1
 93 1adja.dssp1
153 1ahja.dssp1
```

awk 'NR%3' test8.dat



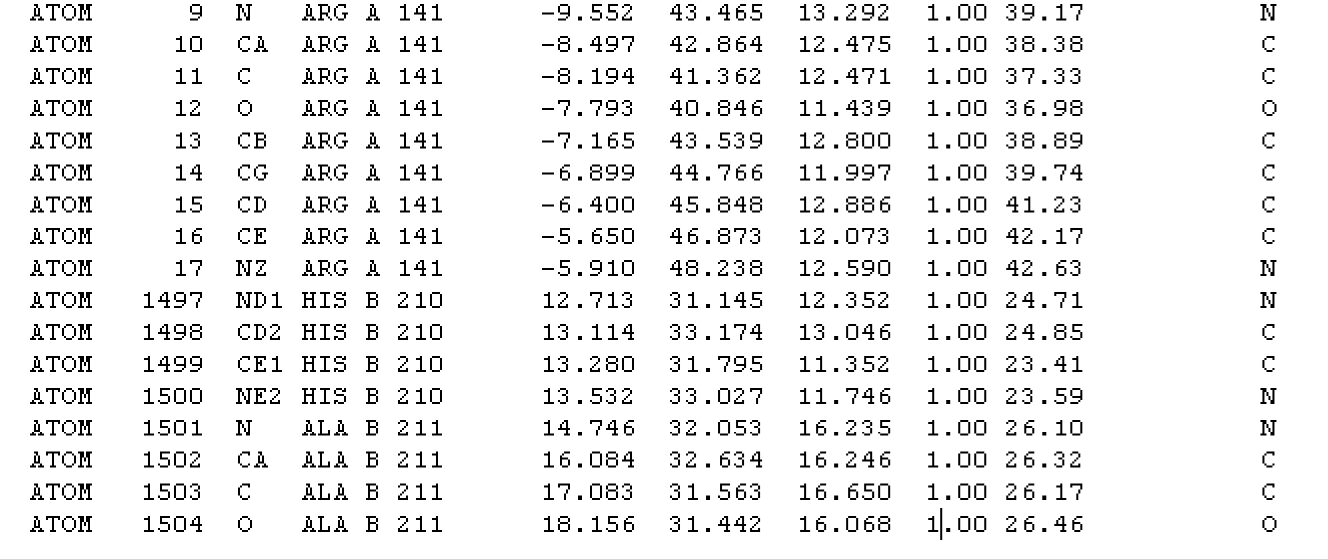
```
139 1a1sa.dssp1
139 1a1sb.dssp1
219 1a2zb.dssp1
350 1a59a.dssp1
385 1a5aa.dssp1
385 1a5ab.dssp1
100 1a6fb.dssp1
103 1a8la.dssp1
 93 1adja.dssp1
 93 1adjb.dssp1
153 1ahjb.dssp1
```



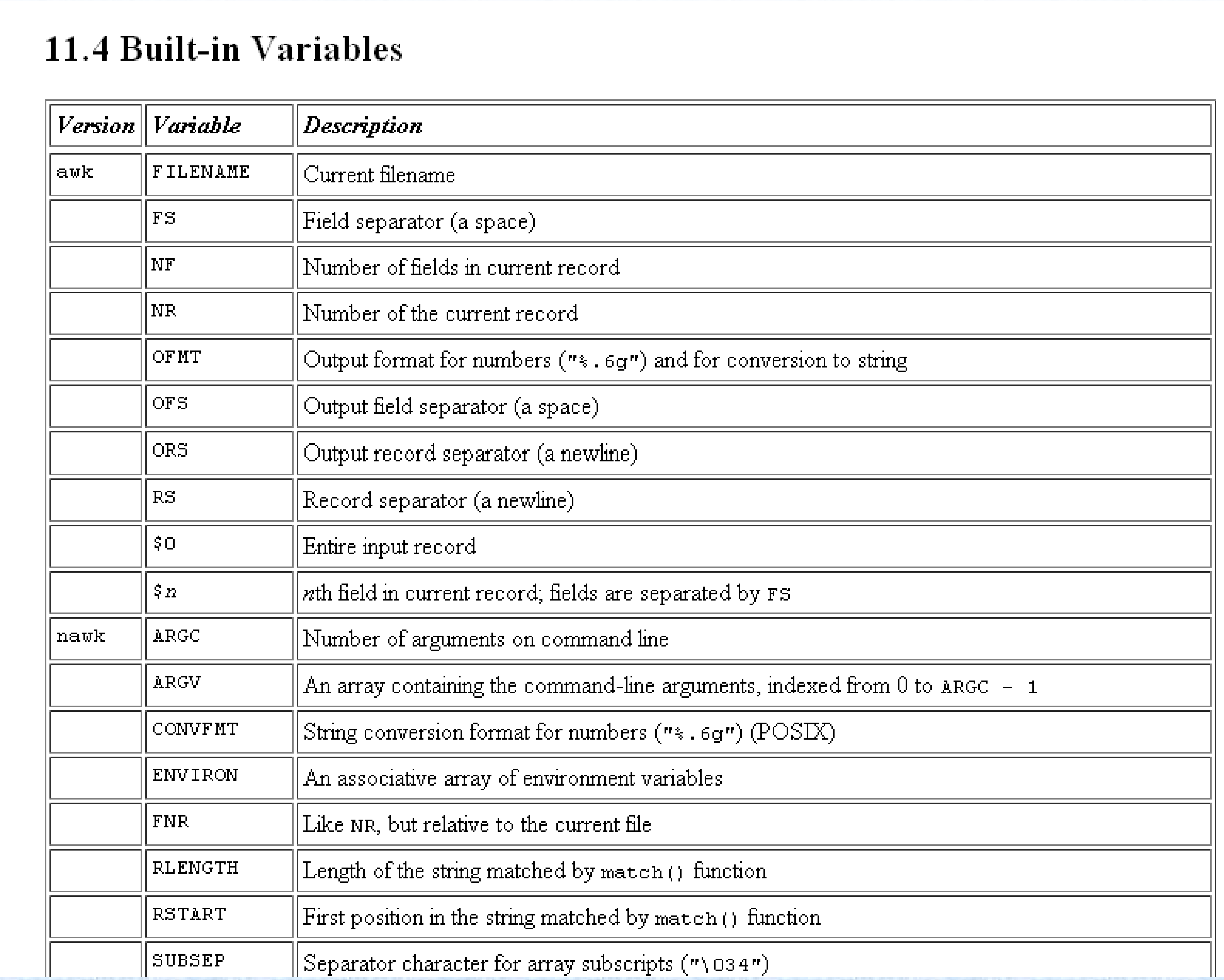
```
139 1a1sa.dssp1
139 1a1sb.dssp1
219 1a2za.dssp1
219 1a2zb.dssp1
350 1a59a.dssp1
350 1a59b.dssp1
385 1a5aa.dssp1
385 1a5ab.dssp1
100 1a6fa.dssp1
100 1a6fb.dssp1
103 1a8la.dssp1
103 1a8lb.dssp1
 93 1adja.dssp1
 93 1adjb.dssp1
153 1ahja.dssp1
153 1ahjb.dssp1
100 1aipa.dssp1
100 1aipb.dssp1
 93 1aipc.dssp1
 93 1aipd.dssp1
341 1aj8a.dssp1
341 1aj8b.dssp1
164 1amua.dssp1
164 1amub.dssp1
 80 1amuc.dssp1
 80 1amud.dssp1
 83 1amue.dssp1
 83 1amuf.dssp1
```

16. Substitute LYS by ARG

awk '{sub(/LYS/,"ARG"); print}' test6.dat



```
ATOM      9  N   ARG A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  ARG A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   ARG A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   ARG A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  ARG A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  ARG A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  ARG A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  ARG A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  ARG A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```



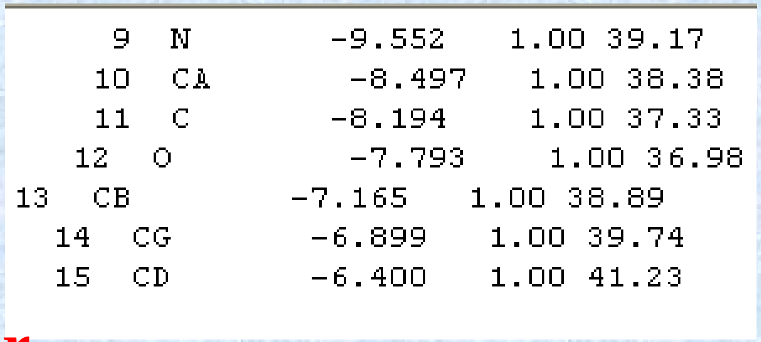
11.4 Built-in Variables

Version	*Variable*	*Description*
awk	`FILENAME`	Current filename
	`FS`	Field separator (a space)
	`NF`	Number of fields in current record
	`NR`	Number of the current record
	`OFMT`	Output format for numbers (`"%.6g"`) and for conversion to string
	`OFS`	Output field separator (a space)
	`ORS`	Output record separator (a newline)
	`RS`	Record separator (a newline)
	`$0`	Entire input record
	`$`*n*	*n*th field in current record; fields are separated by `FS`
nawk	`ARGC`	Number of arguments on command line
	`ARGV`	An array containing the command-line arguments, indexed from 0 to `ARGC - 1`
	`CONVFMT`	String conversion format for numbers (`"%.6g"`) (POSIX)
	`ENVIRON`	An associative array of environment variables
	`FNR`	Like `NR`, but relative to the current file
	`RLENGTH`	Length of the string matched by `match()` function
	`RSTART`	First position in the string matched by `match()` function
	`SUBSEP`	Separator character for array subscripts (`"\034"`)

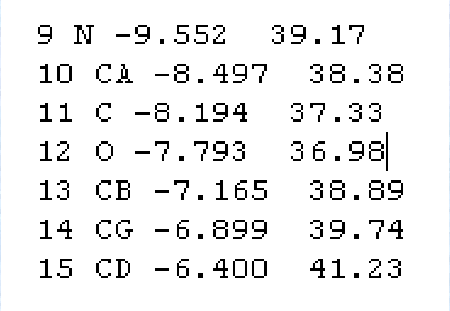
Unix in a nutshell, Stever et al. O'REILLY

17. Delete the 4th field on each line

awk '{$4=""; print}' test9.dat > test10.dat



```
      9   N        -9.552    1.00 39.17
     10   CA        -8.497    1.00 38.38
     11   C        -8.194    1.00 37.33
    12   O        -7.793     1.00 36.98
 13   CB        -7.165    1.00 38.89
    14   CG        -6.899    1.00 39.74
    15   CD        -6.400    1.00 41.23
```

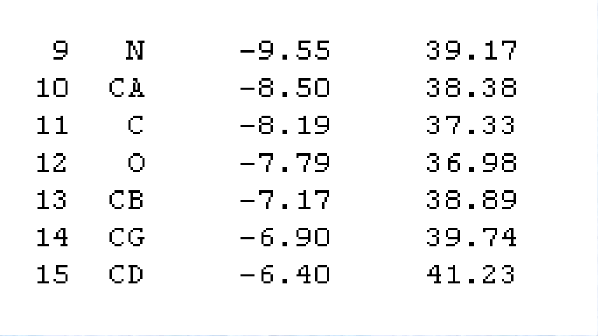


```
9 N -9.552  39.17
10 CA -8.497  38.38
11 C -8.194  37.33
12 O -7.793  36.98
13 CB -7.165  38.89
14 CG -6.899  39.74
15 CD -6.400  41.23
```

18. Place the fields in order

"printf" option

awk '{printf ("%3d %3s %9.2f %9.2f\n", $1,$2,$3,$4)}' test10.dat



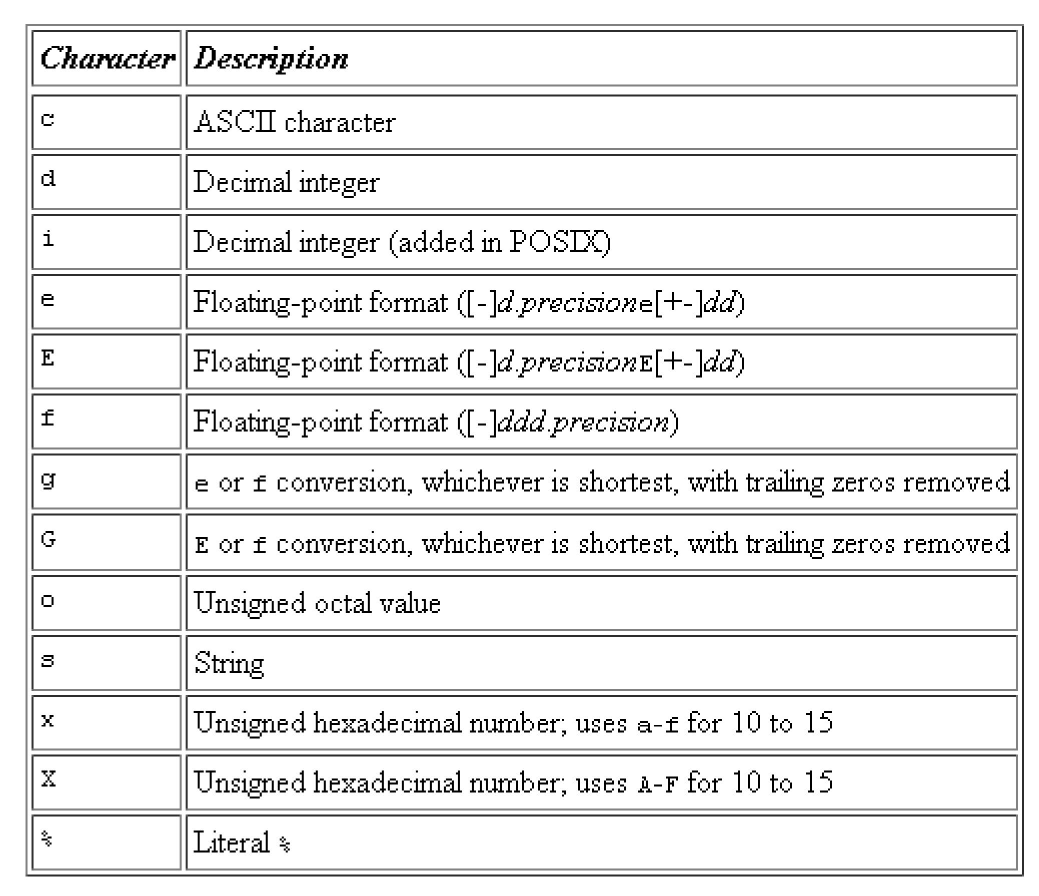
```
  9    N     -9.55      39.17
 10   CA     -8.50      38.38
 11    C     -8.19      37.33
 12    O     -7.79      36.98
 13   CB     -7.17      38.89
 14   CG     -6.90      39.74
 15   CD     -6.40      41.23
```

s: A format (STRING)
d: I format (INTEGER)
f: F format (DECIMAL)

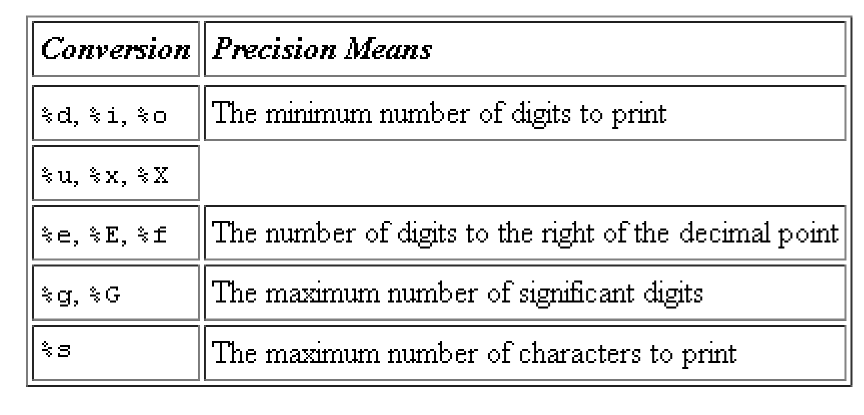
19. Delete 4th field and put place all other fields in order

awk '{$4=""; printf ("%3d %3s %9.2f %9.2f\n", $1,$2,$3,$5)}' test9.dat

Print formats



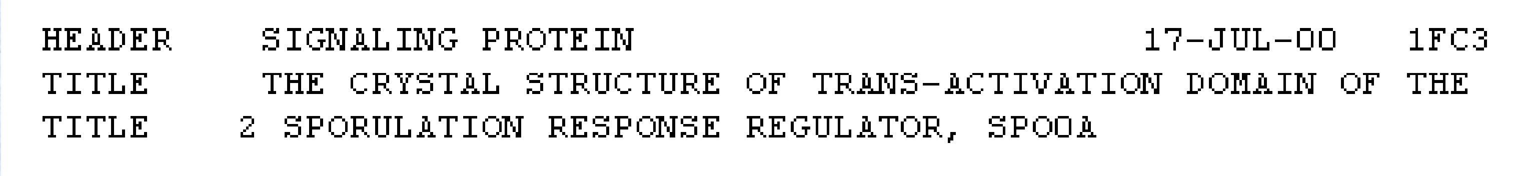
Character	*Description*
`c`	ASCII character
`d`	Decimal integer
`i`	Decimal integer (added in POSIX)
`e`	Floating-point format ([-]*d.precision*`e`[+-]*dd*)
`E`	Floating-point format ([-]*d.precision*`E`[+-]*dd*)
`f`	Floating-point format ([-]*ddd.precision*)
`g`	`e` or `f` conversion, whichever is shortest, with trailing zeros removed
`G`	`E` or `f` conversion, whichever is shortest, with trailing zeros removed
`o`	Unsigned octal value
`s`	String
`x`	Unsigned hexadecimal number; uses `a-f` for 10 to 15
`X`	Unsigned hexadecimal number; uses `A-F` for 10 to 15
`%`	Literal `%`



Conversion	*Precision Means*
`%d`, `%i`, `%o`	The minimum number of digits to print
`%u`, `%x`, `%X`	
`%e`, `%E`, `%f`	The number of digits to the right of the decimal point
`%g`, `%G`	The maximum number of significant digits
`%s`	The maximum number of characters to print

20. Print the first 3 lines

awk 'NR<4 {print}' test2.dat

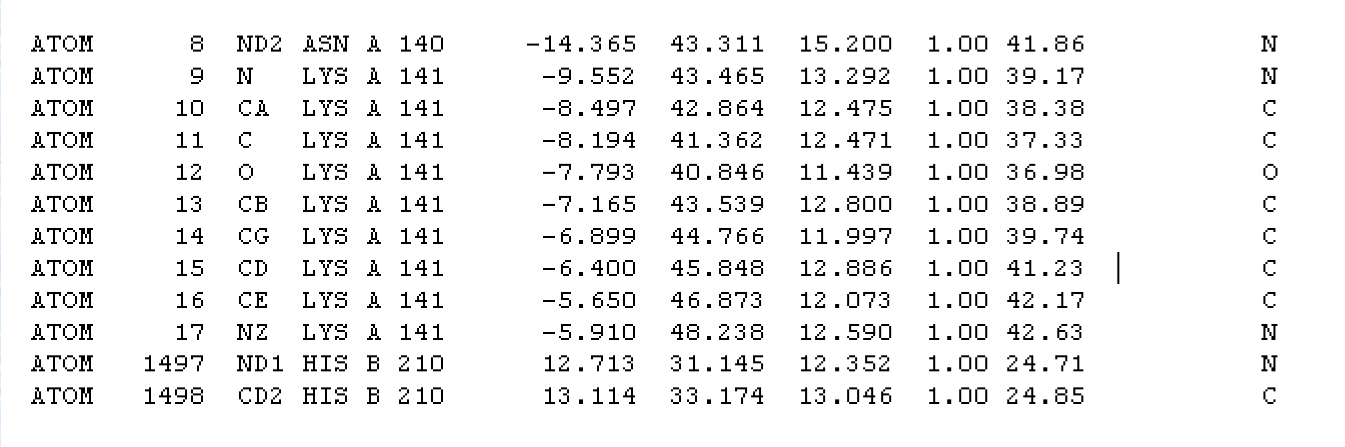


```
HEADER    SIGNALING PROTEIN                                17-JUL-00   1FC3
TITLE     THE CRYSTAL STRUCTURE OF TRANS-ACTIVATION DOMAIN OF THE
TITLE    2 SPORULATION RESPONSE REGULATOR, SPO0A
```

21. Matching strings

Print the lines contains "ATOM"

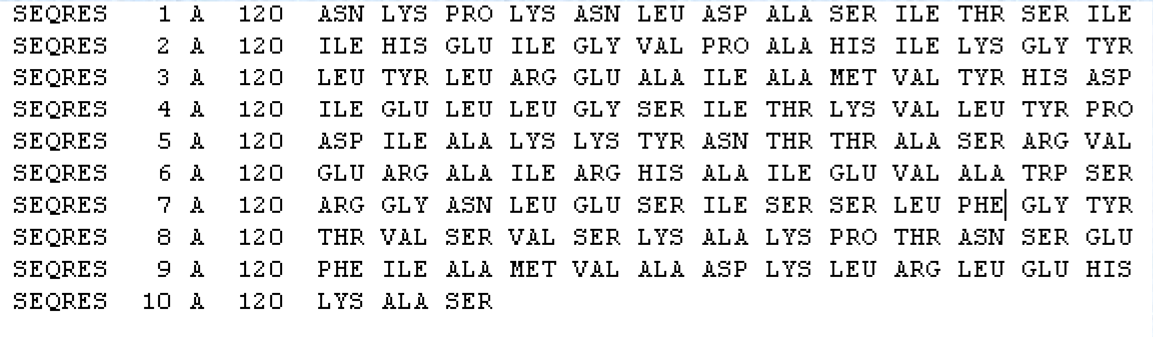
awk '/ATOM/ {print}' test2.dat



```
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86           N
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
```

22. Find the amino acid sequence in a PDB file

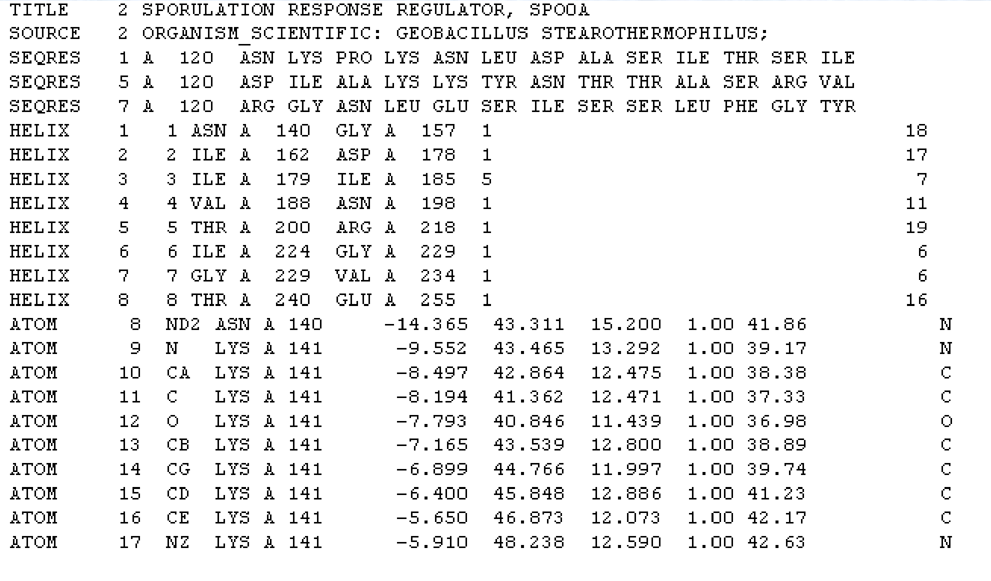
awk '/SEQRES/ {print}' test2.dat



```
SEQRES   1 A  120  ASN LYS PRO LYS ASN LEU ASP ALA SER ILE THR SER ILE
SEQRES   2 A  120  ILE HIS GLU ILE GLY VAL PRO ALA HIS ILE LYS GLY TYR
SEQRES   3 A  120  LEU TYR LEU ARG GLU ALA ILE ALA MET VAL TYR HIS ASP
SEQRES   4 A  120  ILE GLU LEU LEU GLY SER ILE THR LYS VAL LEU TYR PRO
SEQRES   5 A  120  ASP ILE ALA LYS LYS TYR ASN THR THR ALA SER ARG VAL
SEQRES   6 A  120  GLU ARG ALA ILE ARG HIS ALA ILE GLU VAL ALA TRP SER
SEQRES   7 A  120  ARG GLY ASN LEU GLU SER ILE SER SER LEU PHE GLY TYR
SEQRES   8 A  120  THR VAL SER VAL SER LYS ALA LYS PRO THR ASN SER GLU
SEQRES   9 A  120  PHE ILE ALA MET VAL ALA ASP LYS LEU ARG LEU GLU HIS
SEQRES  10 A  120  LYS ALA SER
```

23. Get the atoms of A chain

awk '$5~/A/ {print}' test2.dat

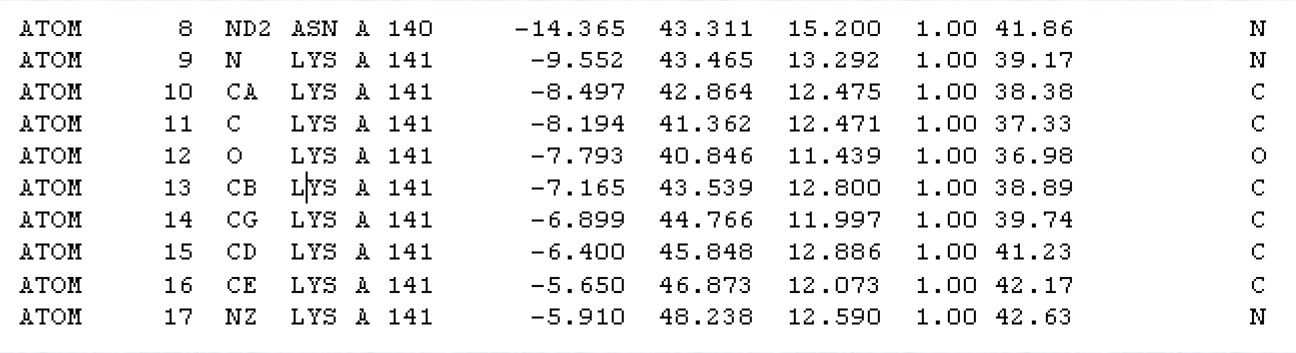


```
TITLE     2 SPORULATION RESPONSE REGULATOR, SPOOA
SOURCE    2 ORGANISM_SCIENTIFIC: GEOBACILLUS STEAROTHERMOPHILUS;
SEQRES    1 A  120  ASN LYS PRO LYS ASN LEU ASP ALA SER ILE THR SER ILE
SEQRES    5 A  120  ASP ILE ALA LYS LYS TYR ASN THR THR ALA SER ARG VAL
SEQRES    7 A  120  ARG GLY ASN LEU GLU SER ILE SER SER LEU PHE GLY TYR
HELIX     1   1 ASN A  140  GLY A  157  1                                  18
HELIX     2   2 ILE A  162  ASP A  178  1                                  17
HELIX     3   3 ILE A  179  ILE A  185  5                                   7
HELIX     4   4 VAL A  188  ASN A  198  1                                  11
HELIX     5   5 THR A  200  ARG A  218  1                                  19
HELIX     6   6 ILE A  224  GLY A  229  1                                   6
HELIX     7   7 GLY A  229  VAL A  234  1                                   6
HELIX     8   8 THR A  240  GLU A  255  1                                  16
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86           N
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
```

Is this the desired one?

24. Use conditions

awk '/ATOM/ && $5~/A/ {print}' test2.dat



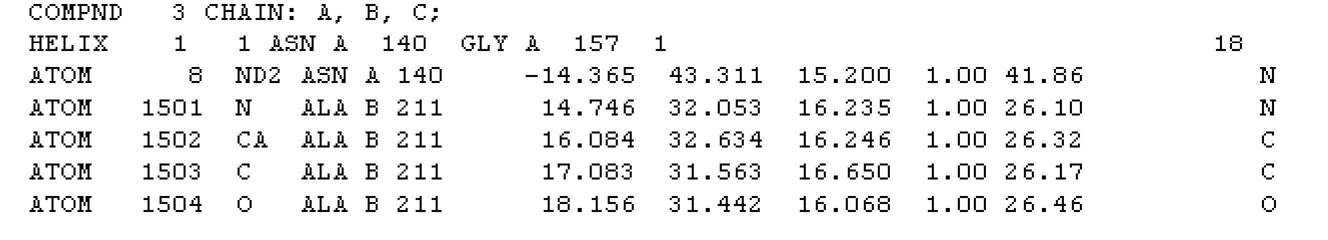
```
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86           N
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
```

25. Strict condition for most probable result

awk '$1~/ATOM/ && $5~/A/ {print}' test2.dat

26. Fourth field starting with A

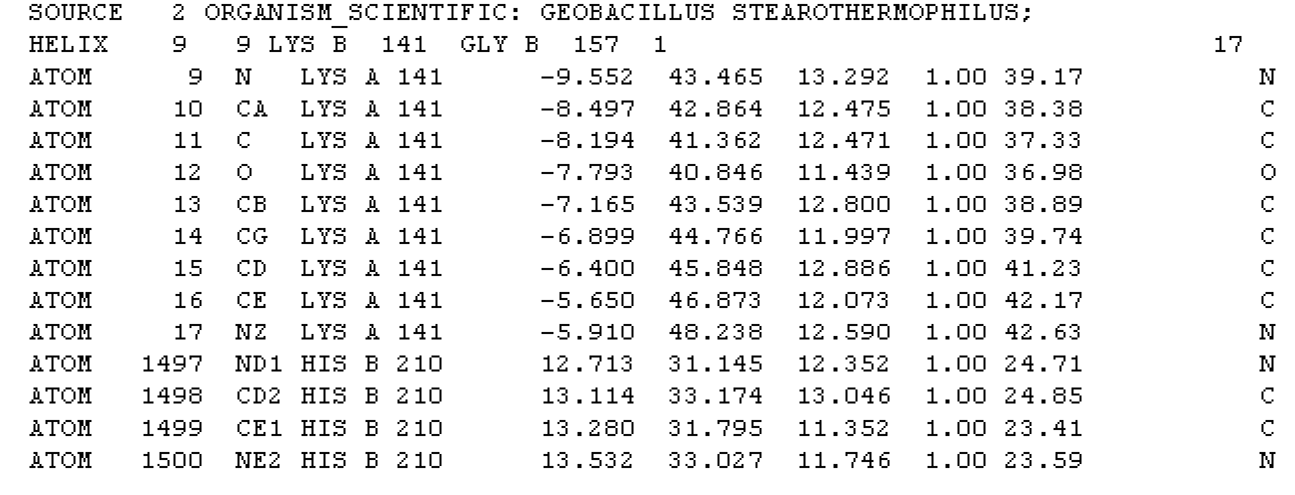
awk '$4~/^A/ {print}' test2.dat



```
COMPND    3 CHAIN: A, B, C;
HELIX    1   1 ASN A  140  GLY A  157  1                                  18
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

27. 4th field ending with "S"

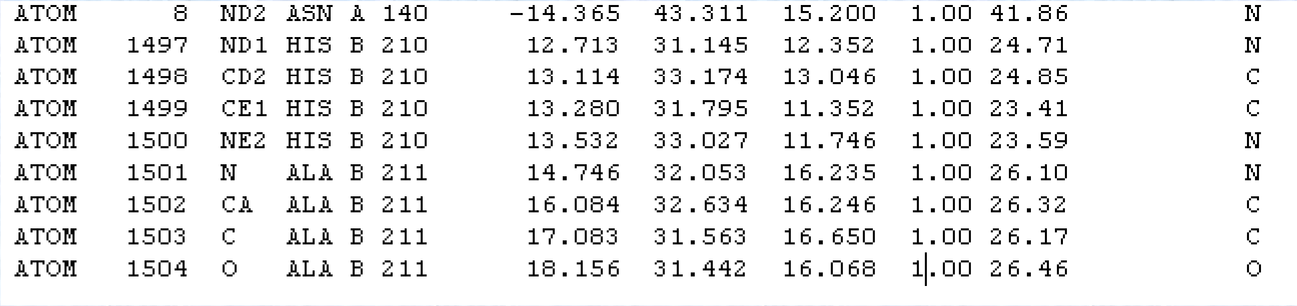
awk '$4~/S$/ {print}' test2.dat



```
SOURCE    2 ORGANISM_SCIENTIFIC: GEOBACILLUS STEAROTHERMOPHILUS;
HELIX    9   9 LYS B  141  GLY B  157  1                                  17
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
```

28. Atoms with no Lys residue

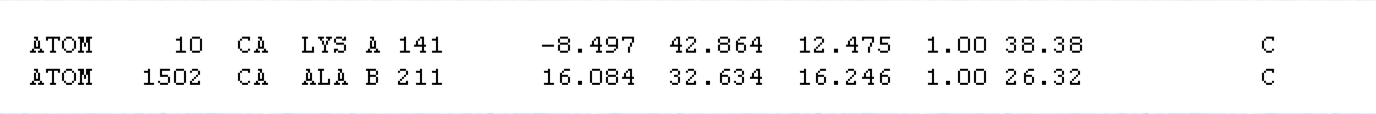
awk '$1~/ATOM/ && $4!~/LYS/ {print}' test2.dat



```
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

29. Get the CA coordinates

awk '$1~/ATOM/ && $3~/CA/ {print}' test2.dat

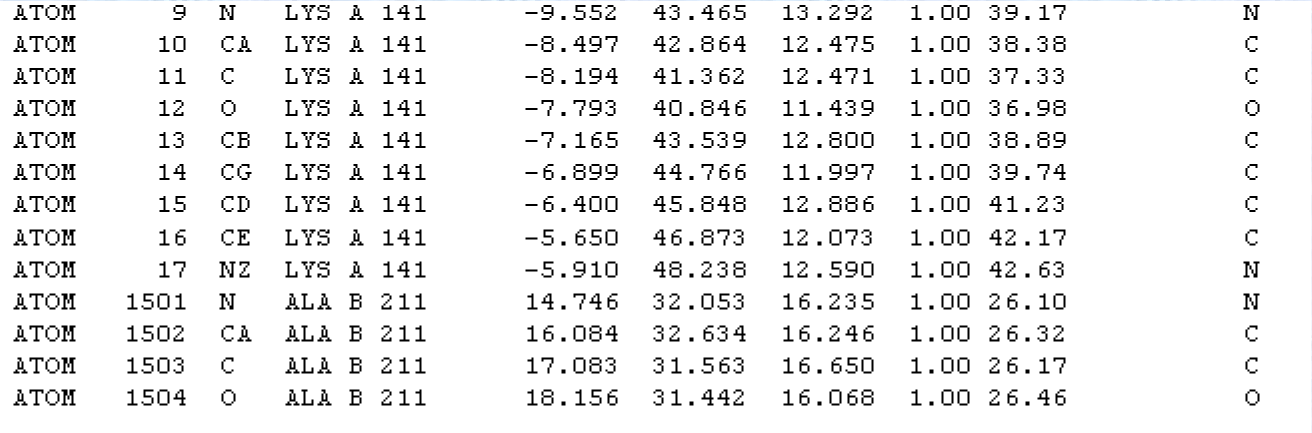


```
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
```

30. Records with LYS or ALA

awk '$1~/ATOM/ && ($4~/LYS/||$4~/ALA/) {print}' test2.dat

awk '$1~/ATOM/ && ($4~/LYS/||/ALA/) {print}' test2.dat



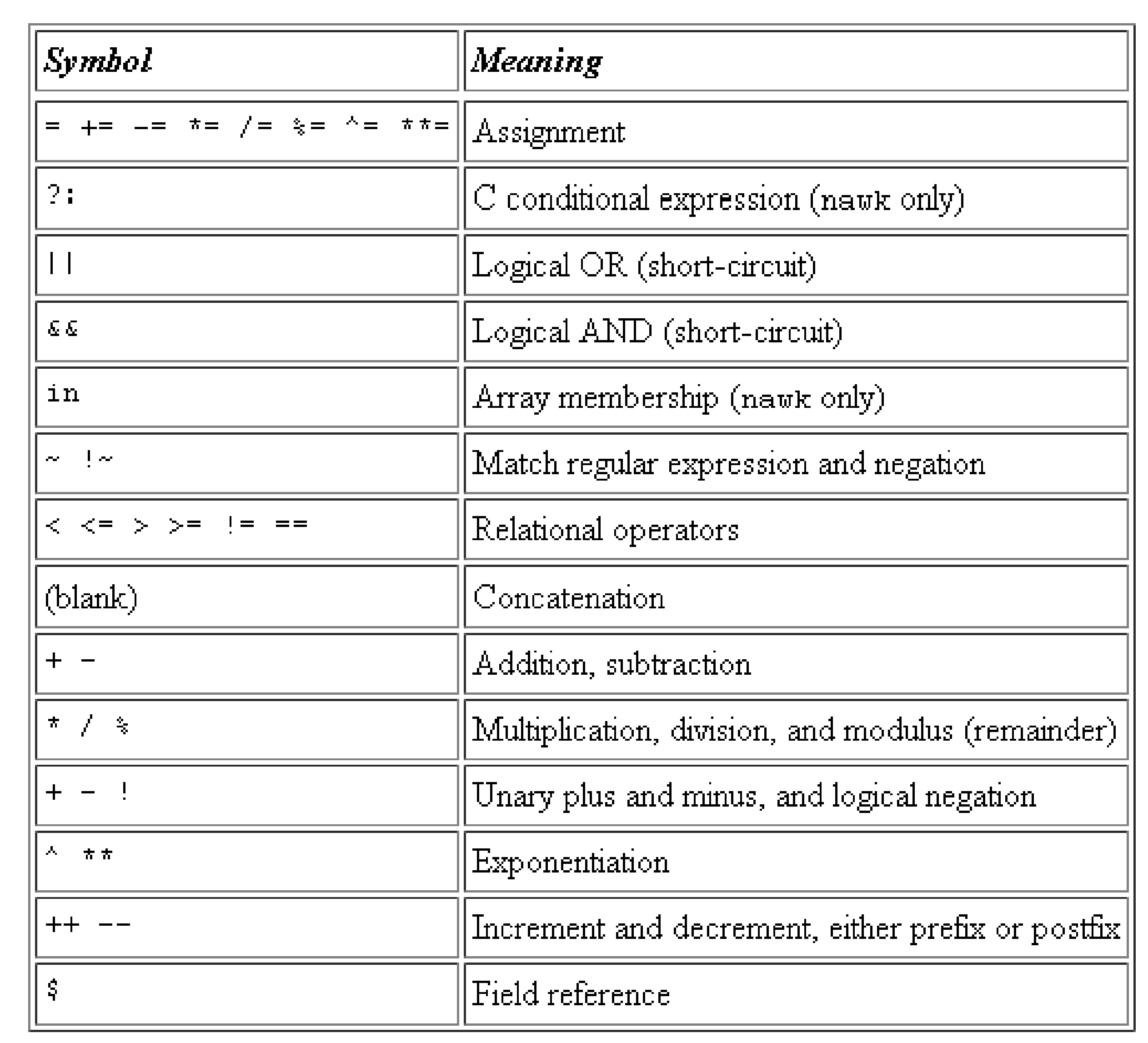
```
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

31. Print the lines that are more than 50 characters

awk 'length > 50' test9.dat

```
12  O        -7.793    1.00 36.98
```

Operators

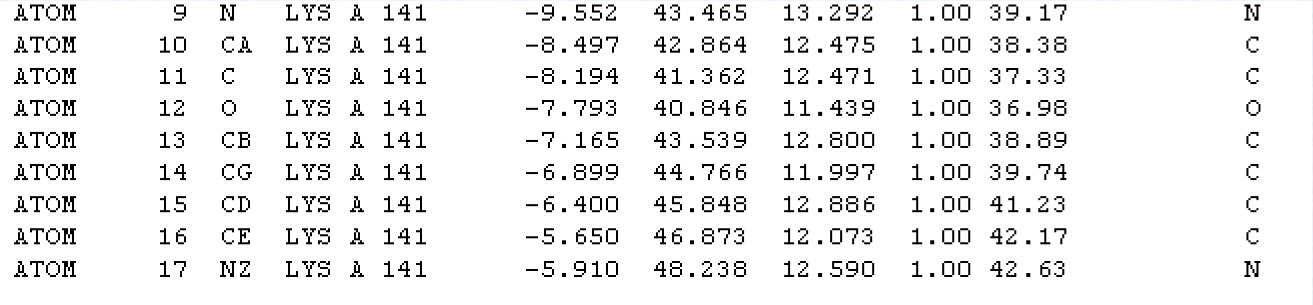


Symbol	*Meaning*
= += -= *= /= %= ^= **=	Assignment
?:	C conditional expression (nawk only)
||	Logical OR (short-circuit)
&&	Logical AND (short-circuit)
in	Array membership (nawk only)
~ !~	Match regular expression and negation
< <= > >= != ==	Relational operators
(blank)	Concatenation
+ -	Addition, subtraction
* / %	Multiplication, division, and modulus (remainder)
+ - !	Unary plus and minus, and logical negation
^ **	Exponentiation
++ --	Increment and decrement, either prefix or postfix
$	Field reference

Unix in a nutshell, Stever et al. O’REILLY

32. Find the atoms with residue number 141

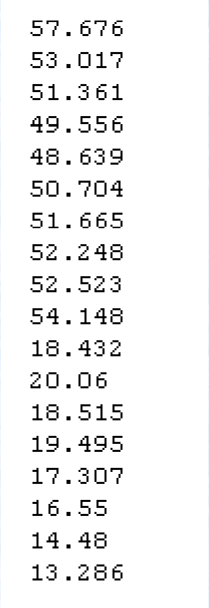
awk '$1~/ATOM/ && $6==141 {print}' test2.dat



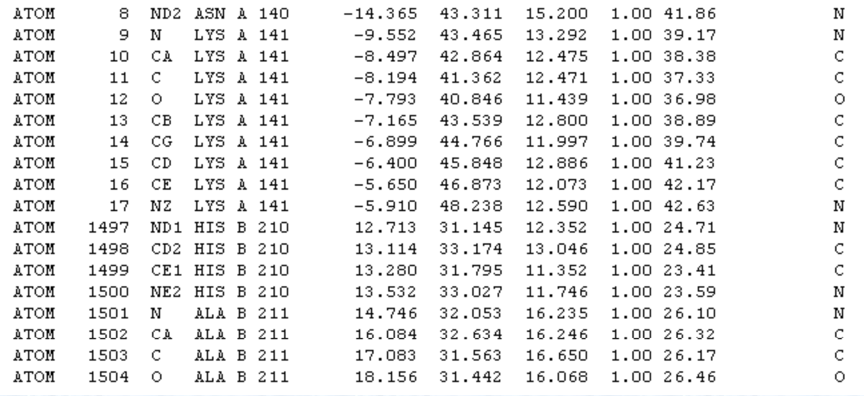
```
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
```

33. Find the difference between X and Y coordinates of all atoms

awk '$1~/ATOM/ {print $8-$7}' test2.dat



```
57.676
53.017
51.361
49.556
48.639
50.704
51.665
52.248
52.523
54.148
18.432
20.06
18.515
19.495
17.307
16.55
14.48
13.286
```



```
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86           N
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

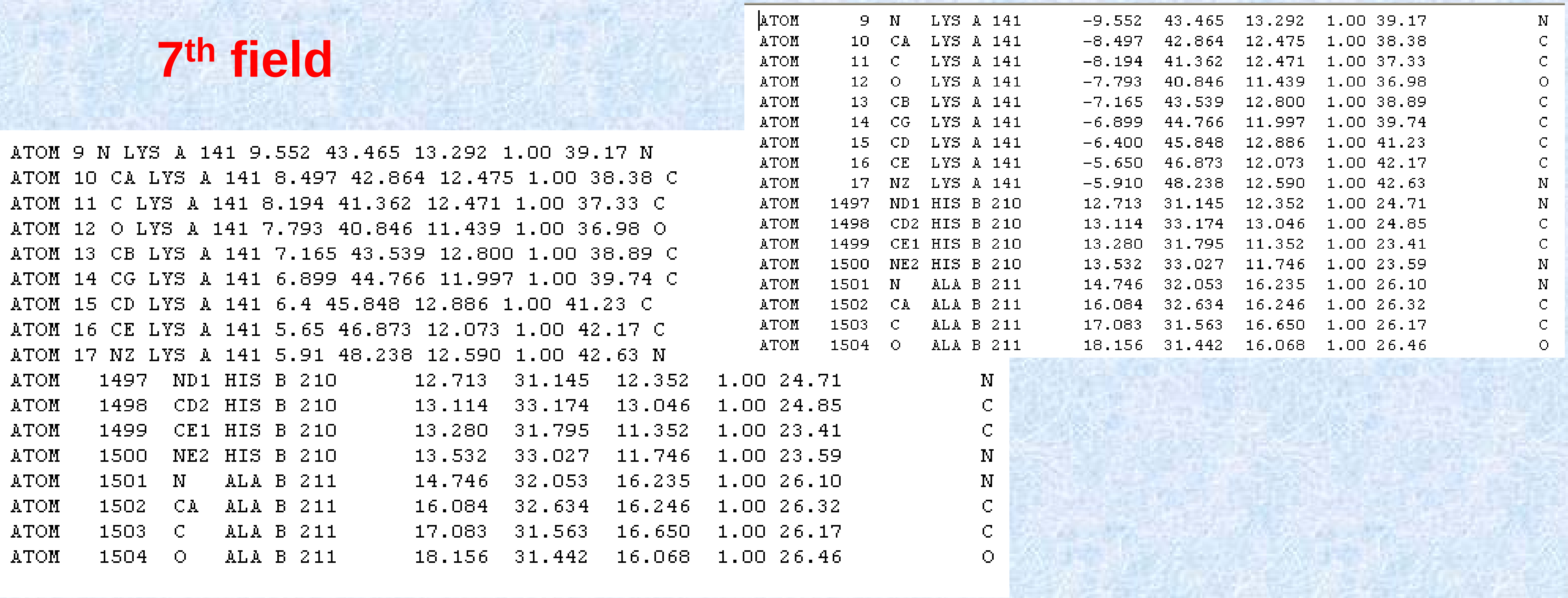
34. Replace by absolute value

```
awk '{for (i=1; i<=NF; i++) if ($i<0) $i=-$i; print}' test6.dat
```

any field

```
awk '{if ($7<0) $7=-$7; print}' test6.dat
```

7th field

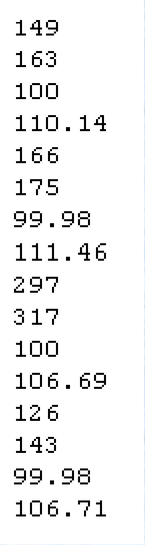


```
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

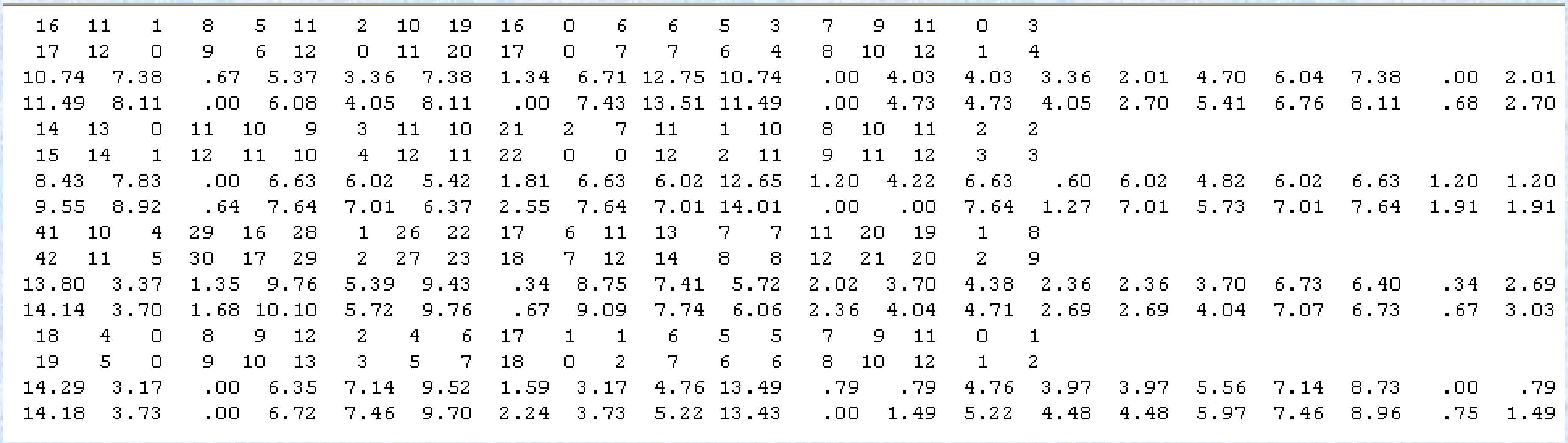
```
ATOM 9 N LYS A 141 9.552 43.465 13.292 1.00 39.17 N
ATOM 10 CA LYS A 141 8.497 42.864 12.475 1.00 38.38 C
ATOM 11 C LYS A 141 8.194 41.362 12.471 1.00 37.33 C
ATOM 12 O LYS A 141 7.793 40.846 11.439 1.00 36.98 O
ATOM 13 CB LYS A 141 7.165 43.539 12.800 1.00 38.89 C
ATOM 14 CG LYS A 141 6.899 44.766 11.997 1.00 39.74 C
ATOM 15 CD LYS A 141 6.4 45.848 12.886 1.00 41.23 C
ATOM 16 CE LYS A 141 5.65 46.873 12.073 1.00 42.17 C
ATOM 17 NZ LYS A 141 5.91 48.238 12.590 1.00 42.63 N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

35. Summing up the numbers in each line

```
awk '{ for(i=1; i<=NF;i++) j+=$i; print j; j=0 }' test5.dat
```



```
149
163
100
110.14
166
175
99.98
111.46
297
317
100
106.69
126
143
99.98
106.71
```

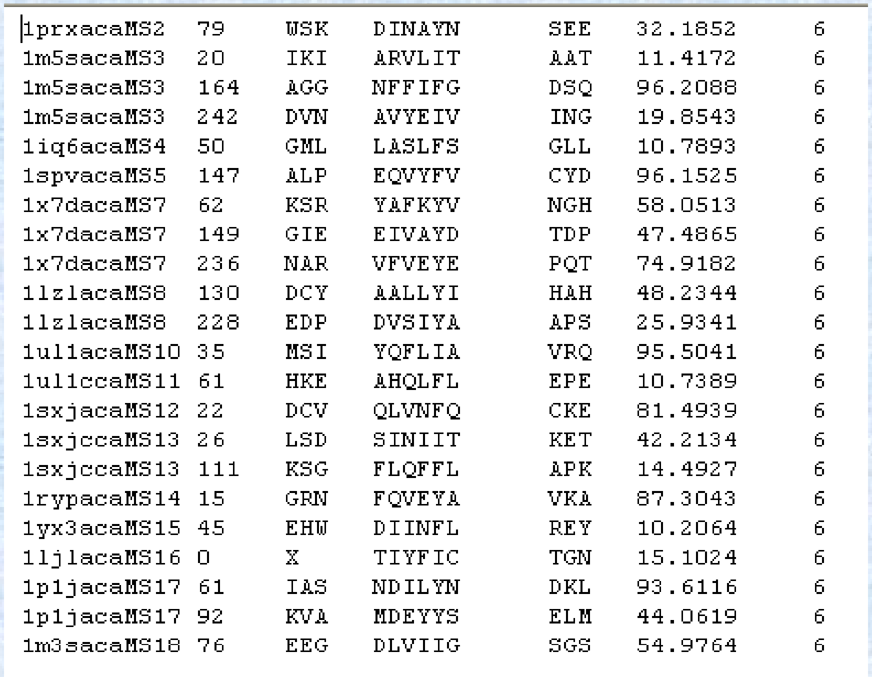


```
    16    11     1     8     5    11     2    10    19    16     0     6     6     5     3     7     9    11     0     3
    17    12     0     9     6    12     0    11    20    17     0     7     7     6     4     8    10    12     1     4
 10.74  7.38   .67  5.37  3.36  7.38  1.34  6.71 12.75 10.74   .00  4.03  4.03  3.36  2.01  4.70  6.04  7.38   .00  2.01
 11.49  8.11   .00  6.08  4.05  8.11   .00  7.43 13.51 11.49   .00  4.73  4.73  4.05  2.70  5.41  6.76  8.11   .68  2.70
    14    13     0    11    10     9     3    11    10    21     2     7    11     1    10     8    10    11     2     2
    15    14     1    12    11    10     4    12    11    22     0     0    12     2    11     9    11    12     3     3
  8.43  7.83   .00  6.63  6.02  5.42  1.81  6.63  6.02 12.65  1.20  4.22  6.63   .60  6.02  4.82  6.02  6.63  1.20  1.20
  9.55  8.92   .64  7.64  7.01  6.37  2.55  7.64  7.01 14.01   .00   .00  7.64  1.27  7.01  5.73  7.01  7.64  1.91  1.91
    41    10     4    29    16    28     1    26    22    17     6    11    13     7     7    11    20    19     1     8
    42    11     5    30    17    29     2    27    23    18     7    12    14     8     8    12    21    20     2     9
 13.80  3.37  1.35  9.76  5.39  9.43   .34  8.75  7.41  5.72  2.02  3.70  4.38  2.36  2.36  3.70  6.73  6.40   .34  2.69
 14.14  3.70  1.68 10.10  5.72  9.76   .67  9.09  7.74  6.06  2.36  4.04  4.71  2.69  2.69  4.04  7.07  6.73   .67  3.03
    18     4     0     8     9    12     2     4     6    17     1     1     6     5     5     7     9    11     0     1
    19     5     0     9    10    13     3     5     7    18     0     2     7     6     6     8    10    12     1     2
 14.29  3.17   .00  6.35  7.14  9.52  1.59  3.17  4.76 13.49   .79   .79  4.76  3.97  3.97  5.56  7.14  8.73   .00   .79
 14.18  3.73   .00  6.72  7.46  9.70  2.24  3.73  5.22 13.43   .00  1.49  5.22  4.48  4.48  5.97  7.46  8.96   .75  1.49
```

36. Summing up all numbers in a particular column

awk '{a+=$7} END {print a}' test4.dat

132



```
1prxacaMS2   79    WSK   DINAYN     SEE   32.1852      6
1m5sacaMS3   20    IKI   ARVLIT     AAT   11.4172      6
1m5sacaMS3   164   AGG   NFFIFG     DSQ   96.2088      6
1m5sacaMS3   242   DVN   AVYEIV     ING   19.8543      6
1iq6acaMS4   50    GML   LASLFS     GLL   10.7893      6
1spvacaMS5   147   ALP   EQVYFV     CYD   96.1525      6
1x7dacaMS7   62    KSR   YAFKYV     NGH   58.0513      6
1x7dacaMS7   149   GIE   EIVAYD     TDP   47.4865      6
1x7dacaMS7   236   NAR   VFVEYE     PQT   74.9182      6
1lzlacaMS8   130   DCY   AALLYI     HAH   48.2344      6
1lzlacaMS8   228   EDP   DVSIYA     APS   25.9341      6
1ul1acaMS10 35     MSI   YQFLIA     VRQ   95.5041      6
1ul1ccaMS11 61     HKE   AHQLFL     EPE   10.7389      6
1sxjacaMS12 22     DCV   QLVNFQ     CKE   81.4939      6
1sxjccaMS13 26     LSD   SINIIT     KET   42.2134      6
1sxjccaMS13 111    KSG   FLQFFL     APK   14.4927      6
1rypacaMS14 15     GRN   FQVEYA     VKA   87.3043      6
1yx3acaMS15 45     EHW   DIINFL     REY   10.2064      6
1ljlacaMS16 0      X     TIYFIC     TGN   15.1024      6
1p1jacaMS17 61     IAS   NDILYN     DKL   93.6116      6
1p1jacaMS17 92     KVA   MDEYYS     ELM   44.0619      6
1m3sacaMS18 76     EEG   DLVIIG     SGS   54.9764      6
```

1. Remove duplicate lines

Consecutive lines

```
awk 'a !~ $0; {a=$0}' test18.dat
```

```
139 1a1sa.dssp1
219 1a2za.dssp1
350 1a59b.dssp1
219 1a2za.dssp1
350 1a59b.dssp1
```

Any lines

```
awk '! a[$0]++ {print}'  test18.dat
```

```
139 1a1sa.dssp1
219 1a2za.dssp1
350 1a59b.dssp1
```

```
sort -u test18.dat
```

test18.dat

```
139 1a1sa.dssp1
139 1a1sa.dssp1
219 1a2za.dssp1
350 1a59b.dssp1
219 1a2za.dssp1
350 1a59b.dssp1
```

Obtain the coordinates of C atom

```
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

```
awk '$1~/ATOM/ && $3=="C" {print}' test6.dat
```

```
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
```

References

Linux in a Nutshell: A desktop quick reference

Siever et al. 2009, Oreilly

http://www.pement.org/awk/awk1line.txt

The AWK Programming Language by Alfred V. Aho, Brian W. Kernighan

Effective Awk Programming, 3/ed, 454 Pages by Arnold Robbins