

Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 28b
“awk” Programming II

(Refer Slide Time: 00:19)

17. Delete the 4th field on each line

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

Handwritten notes: $\text{test9.dat} \rightarrow \text{test10.dat}$

test9.dat	test10.dat
9 N -9.552 1.00 39.17	9 N -9.552 39.17
10 CA -8.497 1.00 38.38	10 CA -8.497 38.38
11 C -8.194 1.00 37.33	11 C -8.194 37.33
12 O -7.793 1.00 36.98	12 O -7.793 36.98
13 CB -7.165 1.00 38.89	13 CB -7.165 38.89
14 CG -6.899 1.00 39.74	14 CG -6.899 39.74
15 CD -6.400 1.00 41.23	15 CD -6.400 41.23

Handwritten notes: 9 N -9.5 39.1
10 CA -8.4 38

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

So, if you want to delete the fourth field; for example, if you see this one everything is 1. So, I don't want this information; I want to eliminate it, so I want to delete the fourth field on each line.

So, in this case you have to put awk and this dollar 4; this is a fourth column that should be empty; don't not write like this. Then write the print test 9 dot dat; this is test 9; so this is the output, this is the test 10. So, here put these invert commas without any field; so it is eliminated, if you see this column is present, this column is present, this column is present; this is the fourth column, this is not present and this is fine here.

So, you can eliminate any column; so in the file. Now if you see this one; so this is not in order, this is arranged in a chaos. So, is it possible to arrange this in order? See this better if you have the specific order for example, it is 9 N minus 9.5 and 39.1. So, 10 should be like this; so CA; this should be exactly 8.4; so, this is 38.1, so if it is to arrange this that look good.

(Refer Slide Time: 01:42)

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	-9.552	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

9.3f
9.1f

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

String %3d → 3 place

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

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

So, if you do this; we can also place a fields in specific order, so we use various formats like the; A format or d format or s format. So, also use option printf; in the earlier cases when you print, which command we used?

Student: Print.

Print just print; so, we do not give any formats, so to give any formats; so, give printf for the format; then the program knows that we are giving some formats. So, now we have to specify the fields; what are the fields you want to write. So, here because this is the one we need; this is the; I want to write 1, 2, 3, 4 this is the four fields you want to write. So, you put dollar 1, dollar 2, dollar 3, dollar 4 and which order you want? What is dollar 1? It is a number, it is a integer.

See integer you can write in d; like in Fortran we use this I format. For example, this percent 3d means this is three places three places in numbers. So, it will give three places; if you see this is 1, 2, 3 gives this order and what is dollar 2? Second field.

Student: String.

Second field it is a string; this is string; so, string we use this s format it is same as A format in Fortran; so, this is s. So, how many fields do you need? If you need 1; you put 1, if you need 2; you need 2. When looking into this PDB; so, we have three sometimes we get OD 1, OD 2, NE 1, NE 2. In this case, if we put dollar 3 S; then you allocate three

fields for three spaces; for this integer. So, 1, 2, 3 this second one is written here and the third one is dollar 3; what is dollar 3?

Student: a number.

Number, this is the

Student: decimal

Decimal; decimal means we have some numbers plus some decimals. Then we need to specify how many digits we require after the decimal point. So, we can use the same f format in the Fortran; so here also its f, 9 point two f. So, what is the meaning of 9? What is the 0.2; point will tells you; mentions how many total fields and totally how many fields after the decimal. So, how many space after the decimal? Two place, so two place after the decimals and totally 9; so 1, 2, 3, 4, 5, 6, 7, 8, 9.

So, if you write 9.3f; what will happen?

Student: one more decimal.

So, will put one more 0; one more 3; if it is 9.1f; now it will cut 1; it will not justify this one. So, it will cut the last digit; I mean do that, so now if you want to write any records in order; you can use this specific formats for this string you use s; for the integer we use d and the decimal we use f and it should match the fields.

If you write it wrong way; for example if you write dollar 2; dollar 1 what will happen? If you write dollar 2 comma dollar one.

Student: Second .

This is the format mismatch because second one this is the string, but we give d; if format mismatch it will give error. So, if you specify any format then you have to make sure that, it should be with the match with this fields. Because this same order, it will take this order; so, the order is also important and you can add in newline; so, this is backslash and n.

So, now the question is; so, we did two steps one is we have the five fields; you wanted to write only the fields 1, 2, 3 and 5 and with these result we arranged in order; is it possible to eliminate this fourth column.

(Refer Slide Time: 06:36)

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

Omit column 4 *format*

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	-9.55	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)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

And write the remaining again in a specific order; so we can combine these two commands and make one common command; awk dollar 4 is equal to inverted commas this means that you need to omit and write everything in printf.

So, first section this will what will happen here? This will omit the column number 4; so omit column 4, then what printf will do?

Student: Format.

Specific format, what is this one? This is a format what we require; then what is this one?

Student: Fields.

This is the field; so, if you see this one and this command can you see any difference? If here 1, 2, 3, 4 because we are taking this 1, 2, 3, 4 continuously. Now here instead of 4; we are writing 5 because so the input file, we are eliminating 4 and this is number 5. So, same format we use; so finally we get the same result.

So, instead of using two commands; this 17 and 18; so in 1 command; you will get all these things.

(Refer Slide Time: 07:47)

Print formats	
Character	Description
c	ASCII character
d	Decimal integer
i	Decimal integer (added in POSIX)
e	Floating-point format (<i>[-]d precision[+ -]dd</i>)
E	Floating-point format (<i>[-]d precisionE[+ -]dd</i>)
f	Floating-point format (<i>[-]ddd precision</i>)
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

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

You can do that; And also we have some different formats for the decimal integer, for the string integer; the decimal integer and so on; so one is the formats you can use. Also you can use some precision also dollar d, dollar i, dollar o; number of digits to print and so on.

(Refer Slide Time: 08:09)

20. Print the first 3 lines

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

PDB

```

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    1 MOLECULE: SPOOA;
COMPND    3 CHAIN: A, B, C;
SOURCE    1 ORGANISM_SCIENTIFIC: GEORACILLUS STEAROTHEROPHILUS;
SOURCE    3 ORGANISM_TAXID: 44121;
SOURCE    4 EXPRESSION_SYSTEM: ESCHERICHIA COLI BL21;
SOURCE    5 EXPRESSION_SYSTEM_TAXID: 511693;
SEQUES    1 A 100 ASD LYS PRO LYS ASN LEU ASP ALA HIS ILE LYS GLY THR
SEQUES    2 A 100 ILE HIS GLY ILE GLY VAL PRO ALA HIS ILE LYS GLY THR
SEQUES    3 A 100 LEU THR LEU ARG GLY ALA ILE ALA MET VAL THR HIS ASP
SEQUES    4 A 100 ILE GLU LEU LEU GLY THR ILE THR LYS VAL LEU THR PRO
SEQUES    5 A 100 ASP ILE ALA LYS LYS THR ASN THR THR ALA SER ARG VAL
SEQUES    6 A 100 GLU ARG ALA ILE ARG HIS ALA ILE GLY VAL ALA THR SER
SEQUES    7 A 100 ARG GLY ASN LEU GLY THR ILE SER SER LEU PHE GLY THR
SEQUES    8 A 100 THR VAL SER VAL SER LYS ALA LYS PRO THR ASN SER GLU
SEQUES    9 A 100 PHE ILE ALA MET VAL ALA ASP LYS LEU ASP LEU GLY HIS
SEQUES    10 A 100 LYS ALA SER
HELIX     1 1 ASN A 140 GLY A 157 1
HELIX     2 2 ILE A 142 ASP A 176 1
HELIX     3 3 ILE A 179 ILE A 185 5
HELIX     4 4 VAL A 188 ASN A 198 1
HELIX     5 5 THR A 200 ARG A 218 1
HELIX     6 6 ILE A 224 GLY A 232 1
HELIX     7 7 GLY A 229 VAL A 234 1
HELIX     8 8 THR A 240 GLY A 255 1
HELIX     9 9 LYS B 141 GLY B 157 1
HELIX    10 10 ILE B 142 ASP B 176 1
HELIX    11 11 ILE B 179 ILE B 185 5
HELIX    12 12 THR B 188 THR B 197 1
HELIX    13 13 THR B 200 GLY B 219 1
HELIX    14 14 THR B 240 LYS B 254 1
ATOM      8 N2O ASN A 140      -0.5145 43.311 15.200 1.00 41.86  N
ATOM      9 N LYS A 141      -0.5532 43.445 15.292 1.00 39.17  N
ATOM     10 CA LYS A 141      -0.4977 42.844 15.470 1.00 38.38  C
ATOM     11 C LYS A 141      -0.184 44.146 15.471 1.00 37.33  C
ATOM     12 O LYS A 141      -7.793 40.846 11.439 1.00 34.98  O
ATOM     13 CB LYS A 141      -7.148 40.539 11.000 1.00 36.89  C
ATOM     14 CG LYS A 141      -6.899 44.744 11.997 1.00 39.74  C
ATOM     15 CD LYS A 141      -6.400 45.840 12.086 1.00 41.23  C
ATOM     16 CE LYS A 141      -5.600 46.870 12.071 1.00 41.17  C
ATOM     17 HE LYS A 141      -5.910 48.238 12.590 1.00 41.43  N
ATOM    1497 HD1 HIS B 210     12.733 16.146 12.182 1.00 24.71  H
ATOM    1498 HD2 HIS B 210     12.124 15.174 12.044 1.00 24.85  H
ATOM    1499 HD3 HIS B 210     12.280 11.799 11.352 1.00 23.41  H
ATOM    1500 HD4 HIS B 210     12.552 15.027 11.744 1.00 23.58  H
ATOM    1501 N ALA B 211     14.744 32.053 16.235 1.00 24.10  N
ATOM    1502 CA ALA B 211     14.084 32.674 16.146 1.00 24.32  C
ATOM    1503 C ALA B 211     17.093 31.561 16.450 1.00 24.17  C
ATOM    1504 O ALA B 211     18.134 31.442 16.048 1.00 24.44  O

```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

So, now the next question is; so whether we can print the first three lines? What is this one; this record? It is a PDB file. So, I want to print the first three lines, how to print this? So, if you NR less than 4; earlier we use the command to print line from 15; that is

more than 14 we printed that. So, if it is less than 4 means it will print only the first three; so this the output that is here.

Now, we come to the important aspects mainly how to manipulate the PDB files because PDB is widely used; we can get various information from the PDB. So, it is important to manipulate the PDB files; further how to get the records or the lines which contains atom; what is the meaning of the atom record?

(Refer Slide Time: 09:01)

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

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

Student: coordinates.

Mainly the coordinates for all atoms; not the hetroatoms; this we want to have the records of the all atoms in a protein or in a DNA. Then we can get this atoms using a command awk, you can see this should contain atom. If it is the backslash means it should contain atom. So, here this is an atom; so it will search the file and wherever we have the atom that will presents. Because I gave print; we did not mention that if 1 or 2 or anything; so, we print the entire line; we can see that.

(Refer Slide Time: 09:44)

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 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 1 SPOILATION RESPONSE REGULATOR, SPOA
SOURCE 1 ORGANISM SCIENTIFIC: OEGACILLUS STEAROTHEROPHILUS
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 238 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 N2 ASN A 140 -14.365 43.311 15.200 1.00 41.86 N
ATOM 9 H LYS A 141 -9.552 43.465 13.292 1.00 39.17 N
ATOM 10 CA LYS A 141 -8.987 42.884 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 11.639 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 44.746 11.997 1.00 39.74 C
ATOM 15 CD LYS A 141 -6.400 45.888 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 46.238 12.590 1.00 42.43 N
```

Is this the desired one?

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

So, now if you are interested to see the sequence; where the sequence information present in the PDB file. How to get the sequence information in PDB file?

Student: SEQRES.

Seqres. This is the column which contains information regarding sequence. So, even though if you want to get this awk; so if you get the sequence where you find the seqres, we are not mentioning which column we need to search, we are just giving the command give the program finds seqres anywhere and print the entire line. If the seqres is present anywhere, it will print; so, here they identify seqres here and print the entire line; so we will get the sequence information.

So, now we go to the another step; so if you want to get the atoms of A chain; where we get the A chain? This is the PDB file, so we need to get the coordinates where shall we get the coordinates?

Student: Atom records.

Atom records where is the chain information? It is here; so fifth field 1, 2, 3, 4, 5; chain is A chain. If you write dollar 5 because dollar 5 is here; 1, 2, 3, 4, 5; dollar 5 contains A and print. Shall we get the A chains atoms? No; we will get A chain along with the additional information. We can see all the; A chain information, this A chain information we will get the records.

But along with we get these additional information; why we get this additional information?

Student: Because those are also contain A.

Here also if you see this A is here 1, 2, 3, 4, 5 this A; 1, 2, 3, 4, 5 this A here; likewise if you see here 1, 2, 3, 4, 5 this A here; here if you take 2, 3, 4, 5 A is here. So, this means if you give this command what will it do?

Student: search field 5

It searches field number 5; whether it contains A? If it contains A, it will print. So, if this some A, this is not the desired one, we do not this; we need only the atom records. So, in this case how can we get the atom records?

Student: first field

So, we can give another condition; the first field should be atom and then fifth field be the A; then we can get this.

(Refer Slide Time: 12:22)

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
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

So, this contains atoms, any record which contains atom and this is the 'and' condition, fifth column starts with A. This is; if you put tilde symbol this starts with A; then you print. So, here we are taking two conditions, last time the fifth column we mentioned this contains A; now we are giving the restriction that it should start with A.

If you give only 5 starts with A; you can eliminate, this you can eliminate, but this no; this starts with A. So, now if you this one; the file starts with A and also it contains atom record, now we check with the atom, everywhere atom and the fifth column is A then you print. But if this record contains atoms somewhere for example, here there is an atom here; then that will print, if this satisfies with A and this is atom that will print.

So, we can strict condition we can give dollar 1; this is atom and the dollar 5 starts with A and then you can print. So, it is very strict condition dollar 1 should be atom; so in this case we will get only atom records. And the fifth column is a there is only change in information; if this is the case definitely you will get the correct result. Then in this case you will get all the atom records without looking in the PDB file; just give this one command, we will get all the atom coordinates for any PDB file

Then the next question is field starting with A; whether we can contains A or we can starts with A or starting with A you can put these symbol.

(Refer Slide Time: 14:19)

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~/SS/ {print}' test2.dat
```

SOURCE	2	ORGANISM SCIENTIFIC: GEORACILLUS STEAROTHERMOPHILUS;						
HELIX	9	N	LYS A 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	NE	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

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

So, A; so here if you see there is test 2; 1, 2, 3, 4, we can see starts with A. So, then we prints, but earlier we see this is the; so, this equivalent to atom. So, the dollar 1 equivalent to atom; this is not the case here, this equal to atom; dollar 1 and dollar 5 is A then we get this. But here this starts with A; this is the difference; one we can say either this contains only A or contains only atom.

Or we can say this contains A or starting with A or ending with A. So, various aspects we can search; here one is we can see with A; or it contains A or starts with A. So, you can see these start with A; all the fifth one which starts with A, you can print. Likewise you can also print ending with some number; some letter which is S. See if this is 1, we need to dollar 4; this ending with this S, you put the S dollar. So, you can see everything this ending with S.

Fourth column 1, 2, 3, 4; write S 1, 2, 3, 4 end with S because this is the one field; so 1, 2, 3, 4. So, I can search with various options either any line contains a character A whatever or it starts with the character or equivalent to the character; equivalent to it is very straight because that is should be only that atom or ends with any of these characters you can give.

(Refer Slide Time: 16:13)

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

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

So, now next question is I need to find the atoms with no LYS and residue. So, if you see this one for example, if you want to see the disulfide bonds whether any protein contains cysteine or not. So, we see the records in this case if you are no LYS and residue. So, here we use exclamatory symbol here; dollar 1 should be atom and dollar 4 should not have LYS. So, if you see 1, 2, 3, 4; if this is the data; LYS means it will omit without LYS it will give all the data. And the print everything means it will print the entire line.

So, if they print the entire line except the lines which contains the lysine. Now another important aspect is when we discussed about the contact maps or the long range order;

we discussed about the C alpha coordinates. We calculate the contact between C alpha atoms; in this case, from PDB file, we need C alpha coordinates. So, we need to extract the C alpha coordinate from the PDB files.

Directly you can get how to get the C alpha coordinates? We can give the conditions; first one dollar 1 should be atom; now this is fine. Then where is C alpha? 1, 2; third column and the third column should be CA; which is there then you can print the entire line; so you will get the C alpha coordinates.

(Refer Slide Time: 17:43)

30. Records with LYS or ALA

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

```
awk '$1~/ATOM/ && ($4~/LYS/||$4~/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	HZ	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

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

Now, here also possible to give either OR; we can see the records is either lysine or with the alanine. So, here we give the information dollar 1 is atom because we need the atom records and dollar 4 because here is one we give the residue name. So, that is, should be lysine, this is the symbol for the OR; this one will give; so I give this dollar 4 is alanine then you print. So, then we can see that this contains with alanine or with lysine.

(Refer Slide Time: 18:24)

Operators

Symbol	Meaning
= += -= *= /= %= ^= ==	Assignment
?:	C conditional expression (awk only)
	Logical OR (short-circuit)
&&	Logical AND (short-circuit)
in	Array membership (awk only)
~ !~	Match <u>regular expression</u> 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

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

So, like this you can see various other symbols like the assignment plus or plus or equal to or the minus or equal to and so on. Like you have the logical operators; so this like the or and you can see this or this is the and. And also you have the regular expression; you can see any of these regular expression you can use.

Relation operators you can use addition, subtraction we can add plus or minus as well as the multiplication division and so on. So, you have various operators available in the awk programming and you can use these operations right for manipulating these files

(Refer Slide Time: 19:09)

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
```

ATOM	8	HDZ	ASN	A	140	-14.365	43.311	15.200	1.00	41.86	N
ATOM	9	H	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	HD1	HS1	B	210	12.713	31.145	12.352	1.00	24.71	N
ATOM	1498	HS2	HS1	B	210	13.114	31.176	13.046	1.00	24.85	C
ATOM	1499	CE1	HS1	B	210	13.280	31.795	11.352	1.00	23.41	C
ATOM	1500	HE2	HS1	B	210	13.532	33.027	11.744	1.00	23.59	N
ATOM	1501	H	ALA	B	211	14.744	32.053	16.235	1.00	16.10	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	16.008	1.00	26.46	O

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

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

I will give you few examples; so, one example is find the atoms with residue number 141 or it is a residue number here.

Student: sixth column.

This is the residue number. So, if you write the dollar one its atom because we need the atom records and dollar 6 1, 2, 3, 4, 5, 6 that is equal to 141 this is of the dollar 6 is equal to 141 then you print. So, it will print all these things right if it is 141 this here then we will print all the records likewise you can do for any for example, if you want to grab particular residue like glycine 10. So, then you put the conditions the atoms. So, the residue names or this column four column will be the glycine, and the sixth column will be the has a 10 you can get the information right if you see get the any specific coordinates for any particular residue.

It is also possible for the operations operators now we can do various operations like plus minus addition subtraction multiplication division right many operations you can do, one example get the difference between two columns for example, 8 and 7. So, how to get the difference. So, we have to just subtract. So, if we want to between 8 and 7 here this see 1, 2, 3, 4, 5, 6, 7, 8. So, if we get the difference between these two columns just we write because this is the coordinates we put atom.

So, print dollar 8 minus dollar 7 otherwise if they say work print dollar 8 minus dollar 7 this will also work in this file see if you take 8 minus 7 what is the result 43 minus 14 plus 14 because you are subtracting. So, 57. So, this is fine. So, these column this 7, this 7 and 8. So, you can have this data for all this coordinates. So, you see get subtraction.

(Refer Slide Time: 21:07)

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	CO	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	24.05	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

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

Ok. So, now, I give a some more examples for example, if you want to replace by absolute values, see we do not want to have this negative values we want this absolute values for any field. So, for i equal to 1 to NF what is NF? Number of fields. So, if you want to do it for any particular field you have to give the particular field any particular field say seventh field if you want to do it for the entire ones. So, then you can give this one right i equal to 1 to the entire field and by plus plus.

If i is less than 0 why it is less than 0.

Student: negative.

Negative values right less than 0 then you replace this i by dollar i by minus dollar i right sorry minus dollar i by dollar i. So, whatever we have the minus values that is replaced by the positive values. So, if you see this one here we have this minus value here right and you can see this minus value is replaced by the plus values.

So, here you can replace any of this values by the absolute value right either you do it for any field or you can do it in particular field. If in particular field you have to mention the field number or any field means you can do it with n two all field. So, you can do it. So, finally, you can replace it.

(Refer Slide Time: 21:28)

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

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

Then also you can do some calculations you can do. So, here I want to sum up the numbers in each line. So, this is the output file. So, this is test 5 dot dat right I want to sum up all these files. So, how to do that? So, i equal to this is the field. So, take from here to all the fields right and then i plus plus is summing up. So, j plus equal to dollar i. So, you are adding the each variables here right and finally, print j. So, print j means this is the sum of the values because we give j plus equal to dollar i. So, j equal to 0 because we need to write this final answer.

So, finally, what will happen to the the first one; you add up all the numbers the number will be 149. So, this line if you see the actual this is the composition of the amino acids right, here this is the output for a specific different amino acids. This is the occurrence, this is the composition of a pair. So, likewise it is going on. See 149 amino acids what is the composition?

Student: Its normalized

Normalized with.

Student: total length

Total length. So, if you add up everything what will happen.

Student: 100.

Should be 100. So, this is the composition. So, if you add up all these numbers these one; that means, that is correct right because if you see the proportion 16 is 10, one is 0.67, 11 is 1.38. So, 16 divided by 149 you get this number. So, finally, if you add up all these numbers right you will get the 100.

So, if you have any files you can write just one simple code with awk and you can do the calculations, you can add you can subtract right you can do anything. So, now, if you want to add all the numbers in a particular column.

(Refer Slide Time: 24:35)

36. Summing up all numbers in a particular column

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

132

hpxacaNS2	79	WSK	DINAYN	SEE	32.1852	6
im5sacaNS3	20	IKI	ARVLIT	AAT	11.4172	6
im5sacaNS3	164	AGG	NFFIFG	DSQ	96.2088	6
im5sacaNS3	242	DYN	APTEIV	IMG	19.8543	6
1iq6acaNS4	50	GML	LASLFS	GLL	10.7893	6
1spvacaNS5	147	ALP	EQVTFV	CYD	96.1525	6
1x7dacaNS7	62	KSR	YAFKTV	NGR	58.0513	6
1x7dacaNS7	149	GIE	EIVATD	TDP	47.4865	6
1x7dacaNS7	236	NAR	VFVEYE	PQT	74.9182	6
1lz1acaNS8	130	DCY	AALLVI	HAR	48.2344	6
1lz1acaNS8	228	EDP	DVSIYA	APS	25.9341	6
1ul1acaNS10	35	HSI	TQFLIA	VRQ	95.5041	6
1ul1acaNS11	61	HSE	AMQFL	FPE	10.7389	6
1sxjacaNS12	22	DCV	QLVNFQ	CKE	81.4939	6
1sxjacaNS13	26	LSD	SINIIT	KET	42.2134	6
1sxjacaNS13	111	KSG	FLQFFL	APK	14.4927	6
1rypacaNS14	15	GRN	FQVEYA	VKA	87.3043	6
1yx3acaNS15	45	EHV	DIINFL	REY	10.2064	6
1lj1acaNS16	0	X	TIYFIC	TGN	15.1024	6
1pijacaNS17	61	IAS	NDILTN	DHL	93.6116	6
1pijacaNS17	92	KVA	NDETYS	ELM	44.0619	6
im3sacaNS18	76	EEG	DLVIIG	SGS	54.9764	6

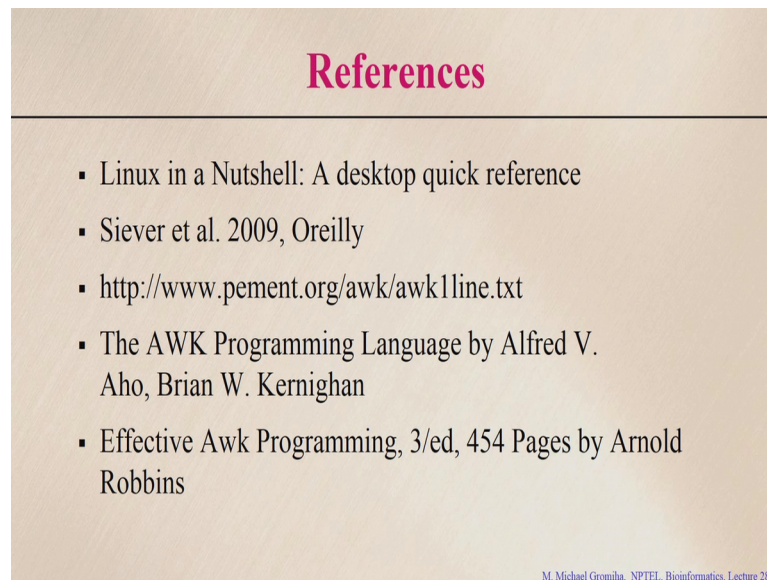
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 28

So, here we added the all the numbers in all the columns we are talking about of this different columns. So, here if you want to add this wise right how to do this. So, here it is column number 1, 2, 3, 4, 5, 6, 7. So, column number seven. So, here we start from the beginning to the end right

Then you print A. So, here how many columns here. So, based on this columns you can add up all this numbers we will get this. So, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22 take this one. So, 22 into 6 132. So, it is a number we get 132 right you can add up all these things.

So, likewise you can either add all the numbers in the column in the new row or you can add different manipulate this columns. So, you can do a lot.

(Refer Slide Time: 25:24)



So, we have got more references; so you can check some of these books Linux in a nutshell or the Oreilly's book and some of these websites. So, you can get more information you can do it one liners, also you can write a program. So, you can combine the different lines; you can write a kind of programming, you can do loops and you can make program to run all these programming in any Linux systems.

So, what we discussed today? Awk programming; awk is the pattern action statement; so what is the main applications of awk?

Student: File Handling.

Mainly file handling specially on the biological data; so, we discussed about the databases and one is servers and the output of the overall files. So, we say bulk of data for example, we take the uniprot that has a large information; lot of information has improved. If you want to extract only particular information; so, either you need to open the file and you have to read properly. It is a very big file, then you need to search the whole file; so, if you write a simple code and get fetch only the information which you require, then easily you can understand the aspects what do you want.

Then for example, in the case of the protein data bank file. So, if you want to work on the C alpha coordinates without looking into these files; just you give one command and you will get the C alpha coordinates and use it for the further programming. Likewise

without editing these files, you can work on these files efficiently you can do and can handle large datasets.

Second aspect is mostly all these databases have uniform format. In this case, it is very easy to use awk programming to handle all these files because they are having the unified format for all these files.

So, we discussed various aspects; further you can do the printing of any specific fields or you can replace the files. Or you can get in the particular coordinates or you can do any arithmetic operations; still many more you can do. So, if you read more details in specific references; so you will get more information and you can use awk, it is very powerful. So, in your calculations and kind of do this scripts for handling the programs and the databases.

So, in the next class I will discuss about some of the important problems in bioinformatics and how to deal with that; mainly the classification problems or the real value problems; like solvent accessibility. Either we can have that two states buried or exposed or you can exactly get the value of this accessible surface area. Likewise, I will discuss about various applications and how to handle these types of problems in bioinformatics.

Thanks for your kind attention.