Bioinformatics course 2

Regular expressions (2)

Virginie Orgogozo

21 Oct 2011

Summary

\w = any letter (A-Z) or digit (0-1) or _

\t = a tab character

\s = a white space character (space, tabs, end-of-line, etc.)

\n or \r = end of line

\d = a digit, from 0 to 9

. = any letter, number or symbol except end-of-line character

Use ( ) to capture part of the text and put it into the replacement term using $1 or \1

+ to match one or more entities

\ = *esc*a*p*e character

Open FPexamples.fta

>CAA58790.1= green fluorescent protein [Aequorea victoria]
MSKGEELFTGVVPILVELDGDVNGQKFSVRGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTFSYGVQCFSRY
PDHMKQHDFLKSAMPEGYVQERTIFYKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKMEYNYNSH
NVYIMGDKPKNGIKVNFKIRHNIKDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSQDPHGKRDHMVLLE
FVTSAGITHGMDELYK
>AAZ67342.1= GFP-like red fluorescent protein [Corynactis californica]
MSLSKQVLPRDVKMRYHMDGCVNGHQFIIEGEGTGKPYEGKKILELRVTKGGPLPFAFDILSSVFTYGNRCFCE
YPEDMPDYFKQSLPEGHSWERTLMFEDGGCGTASAHISLDKNCFVHKSTFHGVNFPANGPVMQKKTLNWEPS
SELITAGDGILKGDVTMFLMLEGGHRLKCQFTTSYKAKKAVKMPPNHIIEHRLVRKEVADAVQIQEHAVAKHFIV
.....

>CAA58790_Aequorea
MSKGEELFTGVVPILVELDGDVNGQKFSVRGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTFSYGVQCFSRY
PDHMKQHDFLKSAMPEGYVQERTIFYKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKMEYNYNSH
NVYIMGDKPKNGIKVNFKIRHNIKDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSQDPHGKRDHMVLLE
FVTSAGITHGMDELYK
>AAZ67342_Corynactis
MSLSKQVLPRDVKMRYHMDGCVNGHQFIIEGEGTGKPYEGKKILELRVTKGGPLPFAFDILSSVFTYGNRCFCE
YPEDMPDYFKQSLPEGHSWERTLMFEDGGCGTASAHISLDKNCFVHKSTFHGVNFPANGPVMQKKTLNWEPS
SELITAGDGILKGDVTMFLMLEGGHRLKCQFTTSYKAKKAVKMPPNHIIEHRLVRKEVADAVQIQEHAVAKHFIV
.....

>CAA58790.1= green fluorescent protein [Aequorea victoria]

(>CAA58790) .1= green fluorescent protein [(Aequorea) victoria]

(>\w+).+\[(\w+) .+ **or** **(\>\w+).+\[(\w+) .+**

$1_$2

>CAA58790_Aequorea

Creating your own wildcard

No wildcard for letters only

[ACGT] = any of these four letters, uppercase

[ATGC]+ = a DNA sequence (with no N)

[A-Z] = any uppercase letter

[A-Za-z] = any letter, lowercase or uppercase

[0-9\.] = any digit or decimal point

(character ranges are based upon the ASCII list)

Open LatLon.txt

```
 21 17'24.68"N
157 51'41.50"W
 38 30'36.62"N
 28 17'16.87"W
  8 59'53.30"S
157 58'13.70"W
 10 24'47.84"N
 51 21'54.61"E
 22 52'41.65"S
 48  9'46.62"E
```

21 17'24.68"N	157 51'41.50"W
38 30'36.62"N	28 17'16.87"W
8 59'53.30"S	157 58'13.70"W
10 24'47.84"N	51 21'54.61"E
22 52'41.65"S	48 9'46.62"E

Open LatLon.txt

```
 21 17'24.68"N
157 51'41.50"W
 38 30'36.62"N
 28 17'16.87"W
  8 59'53.30"S
157 58'13.70"W
 10 24'47.84"N
 51 21'54.61"E
 22 52'41.65"S
 48  9'46.62"E
```

```
21 17'24.68"N   157 51'41.50"W
38 30'36.62"N    28 17'16.87"W
 8 59'53.30"S   157 58'13.70"W
10 24'47.84"N    51 21'54.61"E
22 52'41.65"S    48  9'46.62"E
```

HINT
use [NS]
\r = end of line
\t = tab

Open LatLon.txt

```
 21 17'24.68"N
157 51'41.50"W
 38 30'36.62"N
 28 17'16.87"W
  8 59'53.30"S
157 58'13.70"W
 10 24'47.84"N
 51 21'54.61"E
 22 52'41.65"S
 48  9'46.62"E
```

```
21 17'24.68"N   157 51'41.50"W
38 30'36.62"N    28 17'16.87"W
 8 59'53.30"S   157 58'13.70"W
10 24'47.84"N    51 21'54.61"E
22 52'41.65"S    48  9'46.62"E
```

HINT
use [NS]
\r = end of line
\t = tab

Search for:
"([NS])\r
Replace by :
"$1\t

in jedit

Search for:
"([NS])\r
Replace by :
"\1\t

in Textwrangler

Now replace W and S compas points by a minus value

```
21 17'24.68"N   157 51'41.50"W
38 30'36.62"N    28 17'16.87"W
 8 59'53.30"S   157 58'13.70"W
10 24'47.84"N    51 21'54.61"E
22 52'41.65"S    48  9'46.62"E
```

```
21 17'24.68"N   -157 51'41.50"
38 30'36.62"N    -28 17'16.87"
 -8 59'53.30"   -157 58'13.70"
10 24'47.84"N    51 21'54.61"E
-22 52'41.65"     48  9'46.62"E
```

HINT
use "[WS]

1) match the line
157 51'41.50"W
[0-9]+ [0-9 \.\']+\"[WS]
2) put parenthesis on what is kept
([0-9]+ [0-9 \.\']+\")[WS]
3) replace by :
-$1

in jedit

-\1

in Textwrangler

^ = Negation character

[^A] = anything (letters, punctuation, white space including end-of-line) except A

[^A\r] = anything (letters, punctuation, white space) except A and end-of-line

[^\t]+ = anything except tab character = any data from a table file delimited with tab character
GOOD to pull values from tables
GOOD to switch columns in tables

Open ThalassocalyceData.txt

ConceptName	Depth	Latitude	Longitude	Oxygen	RecordedDate	Salinity	Temperature
Thalassocalyce	348.7	36.71804	-122.0574	1.48	1992-03-02 17:40:09.16	34.134	7.165
Thalassocalyce	520.3	36.749134	-122.03682	0.52	1992-05-05 20:01:10.76	34.234	5.826
Thalassocalyce	118.36	36.83848	-121.96761	1.52	1999-05-14 17:48:27.7	34.045	7.465
Thalassocalyce	200.2	36.723267	-122.05352	1.63	1999-08-09 19:48:42.22	34.024	8.731
Thalassocalyce	100.85	36.726974	-122.04878	2.30	1999-08-09 20:12:11.947	33.916	9.497
Thalassocalyce	1509.6	36.584644	-122.52111	0.95	2000-04-17 20:04:23.663	34.418	2.774

Latitude	Longitude	Depth	Oxygen
36.71804	-122.0574	348.7	1.48
36.749134	-122.03682	520.3	0.52
36.83848	-121.96761	118.36	1.52
36.723267	-122.05352	200.2	1.63
36.726974	-122.04878	100.85	2.30
36.584644	-122.52111	1509.6	0.95

HINT
Use [^\t\n]+

Open ThalassocalyceData.txt

ConceptName	Depth	Latitude	Longitude	Oxygen	RecordedDate	Salinity	Temperature
Thalassocalyce	348.7	36.71804	-122.0574	1.48	1992-03-02 17:40:09.16	34.134	7.165
Thalassocalyce	520.3	36.749134	-122.03682	0.52	1992-05-05 20:01:10.76	34.234	5.826
Thalassocalyce	118.36	36.83848	-121.96761	1.52	1999-05-14 17:48:27.7	34.045	7.465
Thalassocalyce	200.2	36.723267	-122.05352	1.63	1999-08-09 19:48:42.22	34.024	8.731
Thalassocalyce	100.85	36.726974	-122.04878	2.30	1999-08-09 20:12:11.947	33.916	9.497
Thalassocalyce	1509.6	36.584644	-122.52111	0.95	2000-04-17 20:04:23.663	34.418	2.774

Latitude	Longitude	Depth	Oxygen
36.71804	-122.0574	348.7	1.48
36.749134	-122.03682	520.3	0.52
36.83848	-121.96761	118.36	1.52
36.723267	-122.05352	200.2	1.63
36.726974	-122.04878	100.85	2.30
36.584644	-122.52111	1509.6	0.95

HINT
Use [^\t\n]+

([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)

8 times ([^\t\n]+)\t without the last \t

Open ThalassocalyceData.txt

ConceptName	Depth	Latitude	Longitude	Oxygen	RecordedDate	Salinity	Temperature
Thalassocalyce	348.7	36.71804	-122.0574	1.48	1992-03-02 17:40:09.16	34.134	7.165
Thalassocalyce	520.3	36.749134	-122.03682	0.52	1992-05-05 20:01:10.76	34.234	5.826
Thalassocalyce	118.36	36.83848	-121.96761	1.52	1999-05-14 17:48:27.7	34.045	7.465
Thalassocalyce	200.2	36.723267	-122.05352	1.63	1999-08-09 19:48:42.22	34.024	8.731
Thalassocalyce	100.85	36.726974	-122.04878	2.30	1999-08-09 20:12:11.947	33.916	9.497
Thalassocalyce	1509.6	36.584644	-122.52111	0.95	2000-04-17 20:04:23.663	34.418	2.774

Latitude	Longitude	Depth	Oxygen
36.71804	-122.0574	348.7	1.48
36.749134	-122.03682	520.3	0.52
36.83848	-121.96761	118.36	1.52
36.723267	-122.05352	200.2	1.63
36.726974	-122.04878	100.85	2.30
36.584644	-122.52111	1509.6	0.95

HINT
Use [^\t\n]+

([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)\t([^\t\n]+)

8 times ([^\t\n]+)\t without the last \t

replace by :
$3\t$4\t$2\t$5

^ = beginnings, $ = endings

[^A] = anything (letters, punctuation, white space including end-of-line) except A

^A = A at the begining of a line

^ = beginnings, $ = endings

[^A] = anything (letters, punctuation, white space including end-of-line) except A

^A = A at the begining of a line

Latitude	Longitude	Depth	Oxygen
36.71804	-122.0574	348.7	1.48
36.749134	-122.03682	520.3	0.52
36.83848	-121.96761	118.36	1.52
36.723267	-122.05352	200.2	1.63
36.726974	-122.04878	100.85	2.30
36.584644	-122.52111	1509.6	0.95

Search for:
^

Replace by :
Sample\t

?

^ = beginnings, $ = endings

[^A] = anything (letters, punctuation, white space including end-of-line) except A

^A = A at the begining of a line

Latitude	Longitude	Depth	Oxygen
36.71804	-122.0574	348.7	1.48
36.749134	-122.03682	520.3	0.52
36.83848	-121.96761	118.36	1.52
36.723267	-122.05352	200.2	1.63
36.726974	-122.04878	100.85	2.30
36.584644	-122.52111	1509.6	0.95

Search for:
^

Replace by :
Sample\t

?

Sample	Latitude	Longitude	Depth	Oxygen
Sample	36.71804	-122.0574	348.7	1.48
Sample	36.749134	-122.03682	520.3	0.52
Sample	36.83848	-121.96761	118.36	1.52
Sample	36.723267	-122.05352	200.2	1.63
Sample	36.726974	-122.04878	100.85	2.30
Sample	36.584644	-122.52111	1509.6	0.95

```
Mus musculus
Agalma elegans
Frillagalma vityazi
Cordagalma tottoni
```

```
M. musculus
A. elegans
F. vityazi
C. tottoni
```

Search for:
(\w)\w+ (\w+)
Replace by :
$1. $2

```
Mus musculus
Agalma elegans
Frillagalma vityazi
Cordagalma tottoni
```

```
M. musculus
A. elegans
F. vityazi
C. tottoni
```

Search for:
(\w)\w+ (\w+)
Replace by :
$1. $2

```
Mus musculus CY456
Agalma elegans
Frillagalma vityazi X
Cordagalma tottoni
```

```
M. musculus CY456
A. elegans
F. vityazi X
C. tottoni
```

HINT
Use ^

```
Mus musculus
Agalma elegans
Frillagalma vityazi
Cordagalma tottoni
```

```
M. musculus
A. elegans
F. vityazi
C. tottoni
```

Search for:
(\w)\w+ (\w+)
Replace by :
$1. $2

```
Mus musculus CY456
Agalma elegans
Frillagalma vityazi X
Cordagalma tottoni
```

```
M. musculus CY456
A. elegans
F. vityazi X
C. tottoni
```

HINT
Use ^

Search for:
^(\w)\w+
Replace by :
$1\.

+ = one or more, * = zero or more
and other quantifiers

a* zero or more a's
a+ one or more a's
a? zero or one a's (i.e., optional a)
a{m} exactly m a's
a{m,} at least m a's
a{m,n} at least m but at most n a's

+ and * quantifiers match the maximum number of characters

USEFUL :
.* = anything at the end of a line

Open Ch3observations.txt

```
 13 January, 1752 at 13:53  -1.414     5.781     Found in tide pools
17 March, 1961 at 03:46      14   3.6  Thirty specimens observed
1 Oct., 2002 at 18:22    36.51     -3.4221  Genome sequenced to confirm
20 July, 1863 at 12:02   1.74       133 Article in Harper's
```

```
 1752      Jan.  13   13   53   -1.414     5.781
1961       Mar.  17   03   46   14   3.6
2002       Oct.  1    18   22   36.51      -3.4221
1863       Jul.  20   12   02   1.74 133
```

Open Ch3observations.txt

```
 13 January, 1752 at 13:53  -1.414     5.781     Found in tide pools
17 March, 1961 at 03:46      14   3.6  Thirty specimens observed
1 Oct., 2002 at 18:22    36.51     -3.4221   Genome sequenced to confirm
20 July, 1863 at 12:02   1.74       133 Article in Harper's
```

```
 1752      Jan.  13   13   53   -1.414     5.781
1961       Mar. 17   03   46   14   3.6
2002       Oct.  1    18   22   36.51      -3.4221
1863       Jul.  20   12   02   1.74 133
```

1) put parentheses around the parts that will be kept

(13) (Jan)uary, (1752) at (13):(53) (-1.414)   (5.781) Found in tide pools

2) rephrase the query

(\d+) (\w\w\w)[\w\.\,]* (\d+) at (\d+):( )   ( ) ( ) .*

Open Ch3observations.txt

```
 13 January, 1752 at 13:53  -1.414     5.781     Found in tide pools
17 March, 1961 at 03:46      14   3.6  Thirty specimens observed
1 Oct., 2002 at 18:22    36.51     -3.4221   Genome sequenced to confirm
20 July, 1863 at 12:02   1.74       133 Article in Harper's
```

```
 1752      Jan.  13   13   53   -1.414     5.781
1961       Mar. 17   03   46   14   3.6
2002       Oct.  1    18   22   36.51      -3.4221
1863       Jul.  20   12   02   1.74 133
```

1) put parentheses around the parts that will be kept

(13) (Jan)uary, (1752) at (13):(53) (-1.414)   (5.781) Found in tide pools

2) rephrase the query

(\d+) (\w\w\w)[\w\.\,]* (\d+) at (\d+):(\d+) ([-\d\.]+)   ([-\d\.]+)   .*

3) insert white spaces

(\d+)\s+(\w\w\w)[\w\.\,]*\s+(\d+)\s+at\s+(\d+):(\d+)\s+([-\d\.]+)\s+([-\d\.]+) .*

4) write the replacement term

$3\t$2\.\t$1\t$4\t$5\t$6\t$7

For successful operations

Always test your query with Find first

Spell out as many characters as possible
(ex : ^> rather than > in FASTA files)

Try to force the search to match the entire line using $ and ^

Check the number of replacements done

Common operations with regular expressions

Merge or rearrange columns

Join multiple lines into one

Split a name into several elements : nano_128.dat → nano 128

Abbreviate a name

Delete everything on one line before/after XXX