Tutorial3

Shaun Schreiber

23 February 2014

Question 1

Problem

Write a awk script that searches for the word "start" and replace it with "START" .

Implementation

```
function startToSTART ()
{
    gsub(/start/,"START");
    print $0 > FILENAME;
}
{
   startToSTART();
}
```

The command gsub(f,r,[,file]) was used as it searches for the pattern f and replaces it with r. If the file is specified then `$0` is used. The following statement writes it back to the original file.Note nawk has to be used as it allows multiple files to be open at the same time.

Execution

This program is an awk script, but it will only work if the gawk command is used to run it. The directory of where the files are, also needs to be specified.The following terminal command is used to run the script given that it is in the same directory as the Tmp folder."`gawk -f a1.awk ./Tmp/*.*`".

Question 2

Problem

Write an awk script that counts the number of times the word "FINAL" appears in a given file.

Implementation

```
/FINAL/ {x++};
END {
print x;
}
```

The code first checks for the the pattern "FINAL". If the pattern exists it will then execute the x++ statement.

Execution

The following terminal command is used to run the script given that it is in the same directory as the Tmp folder."`gawk -f a2.awk ./Tmp/*.*`".

Question 3

Problem

Write a awk script that search for the pattern "GGTTAA" in the genome-data.txt file. This sequence appears as 6 separately rules in the form "SY=G". If a match is found then the DE rule of the group is displayed.

Implementation

```
BEGIN {
    FS = "//";
    a = 0;
}

{
    a++;
    DELine = "";
    current = "";
    currentlength = "";
    do {
        getline current
        if (match(substr(current,0,2),"DE")) {
            DELine = DELine "" current;
        }

        if (match(substr(current,0,2),"MA") &&
                  match(current,"SY=")      &&
```

```
                !match(current,"*")) {
            split(current,temp3,"'");
            currentlength = currentlength "" temp3[2];
        }
    } while (!match(current,"TAXO-RANGE") &&  a < 480);

    if (index(currentlength,"GGTTAA") != 0) {
        print DELine;
    }
}
END {
}
```

The record delimeter is set to "//" and the field delimeter is set to " `\n` ". For each record the DE rule is stored and the script searches for all of the fields that contains the SY rule. Each field that contains the SY rule is then stripped so that just the value of the SY rule remains. All of these stripped values are concatenated and then matched to the "GGTTAA" pattern. If a match is found the DE rule is printed. Note the 480 in the while condition that is there due to a bug where the number of fields are incorrect.

Execution

The following terminal command is used to run the script. "`gawk -f a3.awk genomedata.txt`".

Question 4

Problem

Execute the following statements and explain why awk reacts differently to them.

- `cat toets.txt | awk '1 {}'`
- `cat toets.txt | awk '0{}'`

Discussion

In both statements there were no output thus there is no visible difference between the results. Only after adding a print statement inside each of the statements a visible difference could be seen. The first statement executed the print statement while the second did not. Thus awk evaluates the statement infront of the open brace. If there is no statement, then by default it evaluates to true and the code inside the brace is executed. If there is a statement then whether it is true or false will determine if the code inside the brace is executed