Master II Oncology
« High-Throughput Technologies in Oncology »

TP : « Introduction to Perl Programming»
or
« The Perfect File's Quest »

Context :

The goal of this tutorial is to compute a file with specific informations needed by your Chief Executive Officer (CEO). He need this file quickly as possible to publish a review that will change human civilization as we know it …

You dispose from several file sources. First of all, you need to analyse this files to get some statistics on the data content. Then you will need to correlate each files with each other to finally produce one filtered file.

To reach this goal, you will need only one weapon : Perl.

You will work on three comma separeted files : patients, genes & samples.

~ Beginner Level ~

Mission 1 : Input & Output

As you recover from a car crash, you forgot all your computer skills. Don't need to go fast, you will need some training first to be in the game again.

A – Write a program that will print out a message.
B – Try now to pass arguments to your program, to print them out into the shell.
C – Define a subroutine in your script to print your message in the shell.
D – Define a subroutine that return a string message into a scalar. Next, print the scalar in the shell.
E – Print something in a new file.
F – Read the messages you wrote in the file you have just created.

Mission 2 : Variables & Control Structures

Take a breath, your memory is coming back slowly.

A – Create an array. Read the content of the array with '*foreach*' control structure.
B – Test condtions on the values of your array. Note : Testing is different if values are numerics or strings.
C – Create an array which contains this values (Revolver,Revolt,Stuff,BlaRevoBla,BlaBlaRevo). Try to write a regex to find the pattern *Revo* in the array. Then try to find occurrence only at the beginning of the string. Finally, only at the end.
D – Create a subroutine to read the content of your array.
E – Create a hash. Read the content : key and values.
F – Create a hash and an array with both numerics values. Some values must be the same between the two variables. Try to read them and to compare each values to see what is the hash key content for equal values.

~ Intermediate Level ~

Mission 3 : Statistics on Patient File

Your chief want you to analyse the result of a clinical study on 200 patients. The file patients.csv is a simplified file downloaded from the TCGA.

A – Read the file and print the ouput.
B – How many patient are classified in each gender ?
C – How many male patient are from a population annotated with 'BLACK OR AFRICAN AMERICAN' ?
D – How many patient from the sub group identified in question C have the sum of Monocytes and Lymphocytes count greater than 20 ?
E – Compute the mean of the 'Age of Diagnosis' for the sub-group identified in C.

Mission 4 : Link Patient File & Sample File

You received an another file from the anatomopathology department that described each sample treated for the patients enrolled in the clinical study.

A – Create a file wich contains the patient id and total of samples for each patient.
To achieve this , you will need to cross the files using a regular expression.
In this case, using a hash could be a good idea...

~ Master Level ~

Mission 5 : Link & Filter with Gene File

Your last file came from the bioinformatic team, it's a dictionnary of all genes. Your boss think there is a potential link between the survival and the GC content of genes involved in cancer for the MALE, BLACK OR AFRICAN AMERICAN population. (Mission 3 – C). He tells you that you should find a low GC content in these genes. You will show him that he's wrong even if you may be fired for that.

A – Create a file with all patient informations filtered by gender and population, with total of samples by patient. Keep '*GC Content*' and '*Gene Biotype*' values also in the final file.
You can reuse the hash you have done at the mission 4, make some nested loops to read and cross the files using specific columns. If you feel on fire, you can compute the '*GC Content*' mean for the genes of this subgroup.
B - What is the '*Gene Biotype*' found for all genes ?
C - What do you notice about '*GC Content*' ?

Congrats you have survived the perfect file's quest but you are fired :-/