Software Project @ Dana-Farber

Introduction

As part of our application process, we ask that each applicant write a complete Java application. You are free to design the program any way you like, but we require that it be written in Java, and that it be fully documented.

Background

For this project, you are to create a command line Java tool that retrieves cancer genomic data from a remote web service, and summarizes the results.

Your program will need to access cancer genomic data from the cBio Cancer Genomics Portal, which currently supports a REST-based web API.

For example, the following URL:
http://www.cbioportal.org/webservice.do?cmd=getProfileData&genetic_profile_id=gbm_tcga_mutations&id_type=gene_symbol&gene_list=TP53&case_set_id=gbm_tcga_cnaseq

will retrieve all mutations for the gene TP53 in Glioblastoma patients assessed as part of The Cancer Genome Atlas (TCGA) project.

Likewise, the following URL will retrieve all copy number alterations for TP53 in the same set of Glioblastoma patients:
http://www.cbioportal.org/webservice.do?cmd=getProfileData&genetic_profile_id=gbm_tcga_gistic&id_type=gene_symbol&gene_list=TP53&case_set_id=gbm_tcga_cnaseq

Each of these calls will return a simple tab-delimited output: one gene call per patient ID.

For mutation data:

- 0 or NaN = no mutation
- any other string = mutation, e.g. V216M indicates a mutation.

For the copy number data:

- 0 = no change
- NA = Data not available
- -1 or +1 = single copy of gene is lost or gained (you can ignore these)
- -2 = both copies of the gene are deleted
- +2 = multiple copies of the gene are observed

Command Line Tool

Your tool should summarize genomic alterations for the same set of TCGA GBM patients described above.

For example, in the simplest instance, a user would execute your program with a single gene, and output a simple summary.

```
./gbm_summarize.sh TP53
TP53 is mutated in 29% of all cases.
TP53 is copy number altered in 2% of all cases.

Total % of cases where TP53 is altered by either mutation or copy
number alteration: 30% of all cases.
```

However, the user should also be able to execute your command line program with up to three genes. For example:

```
./gbm_summarize.sh TP53 MDM2 MDM4
TP53 is altered in 30% of cases.
MDM2 is altered in 10% of cases.
MDM4 is altered in 10% of cases.

The gene set is altered in 47% of all cases.
```

If you want to check you answers, you can try the cBio Cancer Genomics Portal, which provides a visual front-end to the same data: http://cbioportal.org.

Sending your program

When you are done with your program, please send it as a tar.gz file to Ethan Cerami. Please also include a simple README file with instructions on compiling / running your program.