Software Project @ Dana-Farber

Command line

Introduction

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

Background

For this project, you are to create a command line 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. Interactive documentation is located here: http://www.cbioportal.org/api/swagger-ui.html

For example, the following request:

```
curl -X POST
"http://www.cbioportal.org/api/molecular-profiles/gbm_tcga_mutations/mutations/fetch?projection=S
UMMARY&pageSize=10000000&pageNumber=0&direction=ASC" -H "accept: application/json" -H
"Content-Type: application/json" -d "{ \"entrezGeneIds\": [ 7157 ], \"sampleListId\":
\"gbm_tcga_cnaseq\"}"
```

will retrieve all mutations for the gene TP53 in Glioblastoma patients assessed as part of The Cancer Genome Atlas (TCGA) project. The parameter projection=META will cause the API to return summary statistics in the header of the response rather than the actual data. If a projection is not specified then a list of json objects will be returned, each object contains information about the mutation call for a sample in the specified project, notability the “mutationType” field will indicate what type of mutation the object is describing.

Feel free to leave the study and sample list parameters hardcoded to the glioblastoma study (gbm_tcga). The only parameter you’ll need to work with is entrezGeneId.

entrezGeneIds
The Entrez gene id can be obtained from the included file “gene_results.1000.tsv” which contains the symbol, and numerical id of the top 1000 genes from the NCBI. It is recommended that this file be loaded first and then the command line tool can map between the text symbol and numerical identifier.

Likewise, the following URL will retrieve all copy number alterations for TP53 in the same set of Glioblastoma patients:

```
curl -X POST
"http://www.cbioportal.org/api/molecular-profiles/gbm_tcga_gistic/discrete-copy-number/fetch?discreteCopyNumberEventType=ALL&projection=SUMMARY" -H "accept: application/json" -H
"Content-Type: application/json" -d "{ \"entrezGeneIds\": [ 7157 ], \"sampleListId\":
\"gbm_tcga_cnaseq\"}"
```

The resultant object will be a list of json objects which describe each discrete copy number call event. The “alteration” field of each object will be:

- 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 your contact. Please also include a simple README file with instructions on compiling / running your program.