Technical test for Data Scientist (DS) position in the Computational Biology team

Multiple Myeloma data challenge

Clinical context

Multiple Myeloma (MM) is a type of bone marrow cancer. Treatment for MM involves combinations of drugs over multiple cycles. There is huge heterogeneity in treatment response with some individuals being non-responders and some patients remaining well for some time before a relapse. A better characterization of patients who relapse early can influence the treatment options and combinations.
In this test, we propose to develop a model for predicting the risk of dying or relapsing of newly diagnosed multiple myeloma patients from baseline clinical and expression data.

Data

The data for this test are extracted from an old Synapse Dream Challenge (https://www.synapse.org/#!Synapse:syn6187098/wiki/401884) .
It consists of clinical data, gene expression data and follow-up for newly diagnosed Multiple Myeloma patients extracted from the MMRF CoMMpass IA9 study. In the data, newly diagnosed MM patients are classified as High Risk (HR) when they relapse or die before 18 months.

To access the data, you first need to create an account and download the following files:

- Expression data:
  MMRF_CoMMpass_IA9_E74GTF_Salmon_entrezID_TPM_hg19.csv
  (https://www.synapse.org/#!Synapse:syn10573789)
  [notice that the first column gives Entrez IDs for genes]

- Clinical data and labels:
  sc3_Training_ClinAnnotations.csv
  (https://www.synapse.org/#!Synapse:syn9926878)

- Explanation of the clinical and label annotations:
  Harmonized_Clinical_Dictionary
  (https://www.synapse.org/#!Synapse:syn9744732)

Goal

The purpose of this technical test is to develop a model for predicting the risk of fast dying or relapsing of newly diagnosed multiple myeloma patients (using the High Risk label HR_FLAG).

The evaluation will mostly rely on the way you approach the problem: pre-analysis, preprocessing strategy, choice of modelization and coding skills.
The code should be developed so that the model can be applied to an external validation dataset. You will send your code (Notebook or script) along with a small report to interpret the model and put it in MM context (the use of the literature is clearly welcome).

Your model can be developed in Python or R with a small README to explain how to apply it to external data.
You can use external knowledge/data to develop the model. Please add all the requirements for libraries that should be installed to make it run.

If not used to survival analysis, the candidate can consider a simplified version in which it can assumed that no censored patients will be present in the external validation dataset.

(Obvious) suggestion: OS and PFS related variables are also labels and not features: HR_FLAG is defined as OS or PFS < 18 months (taking into account censoring).