Первое домашнее задание по биоинформатике

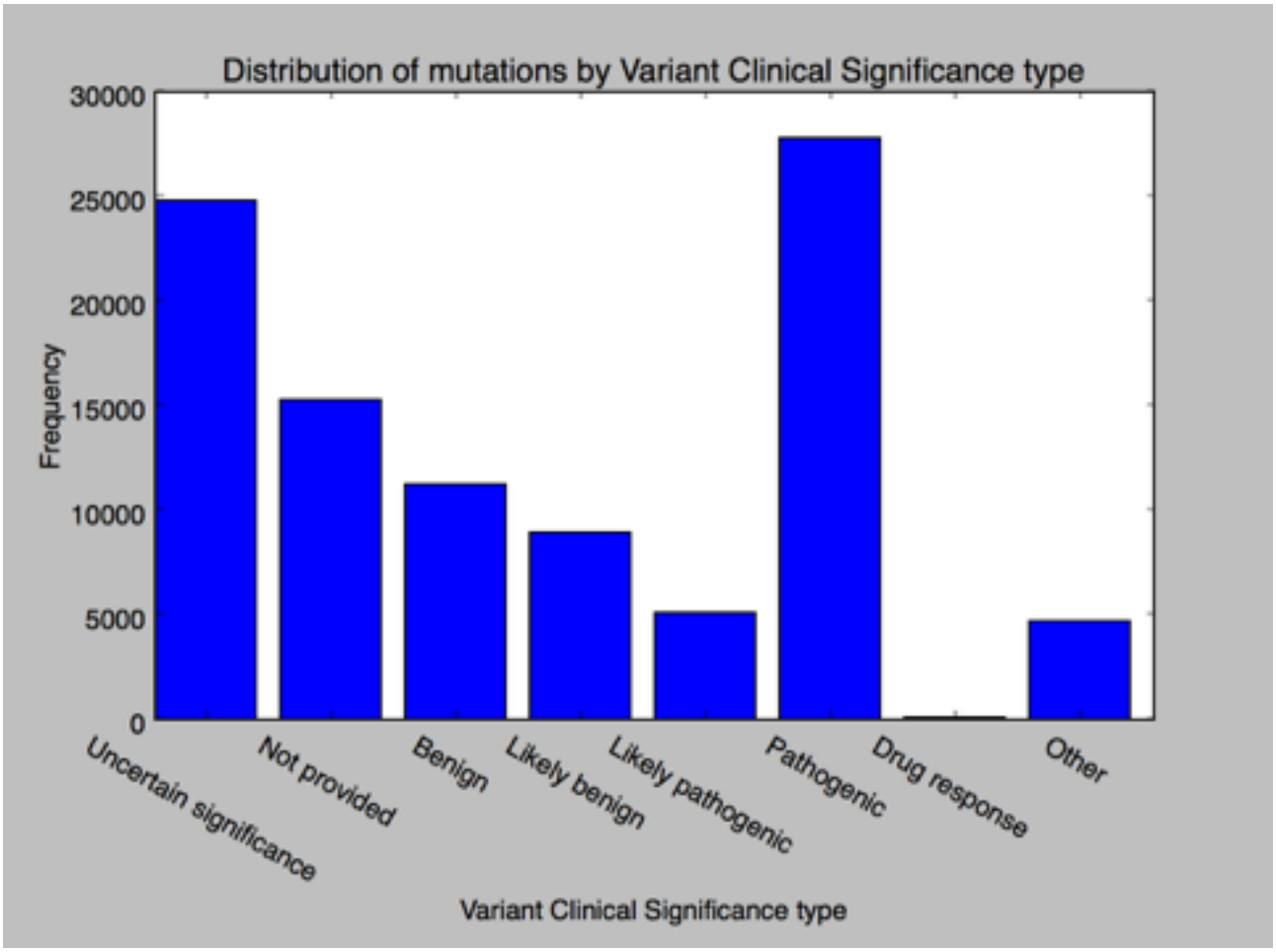
В скрипт я вводил имя “EugeneYurtaev”. Мой номер домашнего задания - 4.

Само задание:

База clinvar
Необходимо скачать базу Clinvar разных лет ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive/. Это база в формате VCF, описывающая разные мутации в геноме человек.
Требуется посчитать распределение мутаций по полю Variant Clinical Significance (его возможные значения описаны в заголовке базы) и построить гистограмму по этому полю. Также требуется построить графики роста количества мутаций в базе (отдельный кривую по каждому значению Variant Clinical Significance и все их на одном графике).

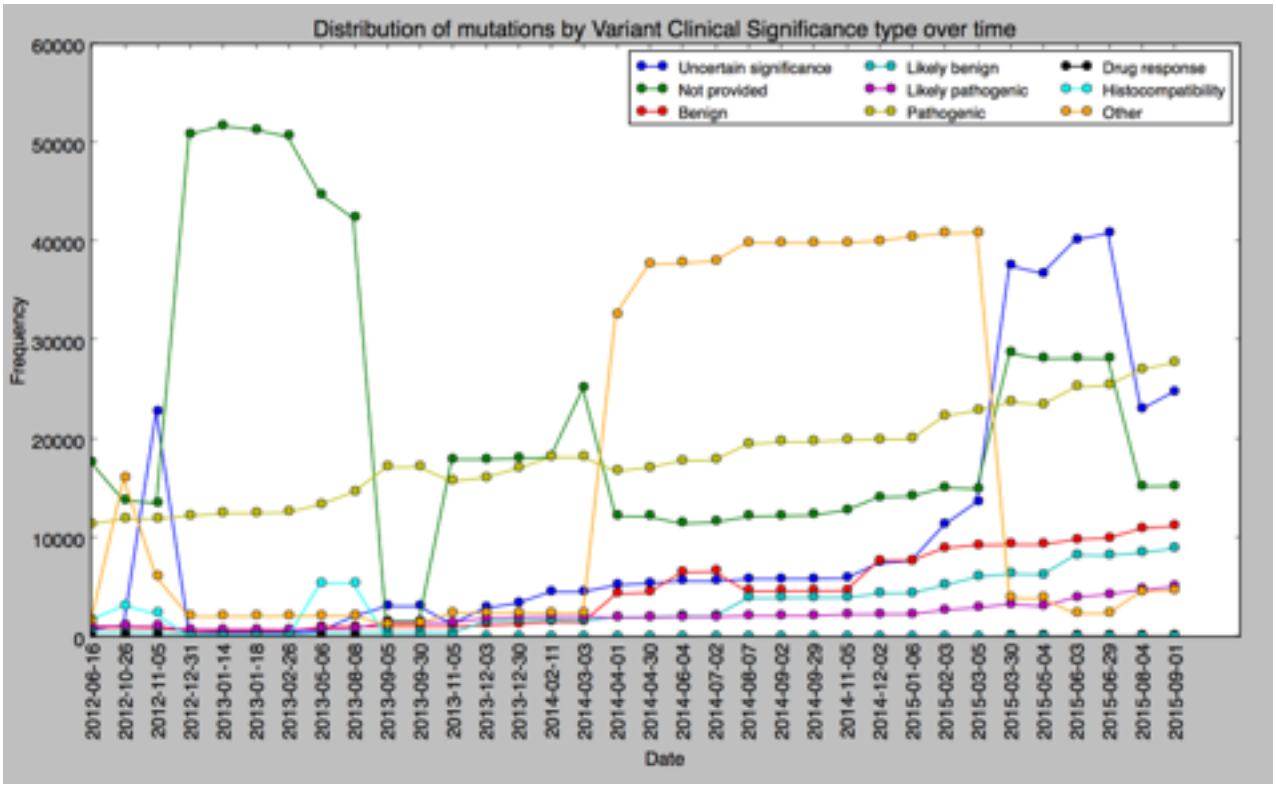
Чтобы скачать все данные с сервера, я использовал консольную утилиту ftp.

Для того, чтобы посчитать распределение мутаций по полю Variant Clinical Significance, я взял последние данные и написал скрипт на питоне, который находит значение поля CLNSIG для каждой мутации и подсчитывает их число.



Так выглядит получившаяся гистограмма.

Второе задание заключилось в том, чтобы посмотреть, как менялась база знаний о мутациях во времени. Для этого я подсчитал количество мутаций по типам в каждом файле и построил графики каждой мутации в отдельности.



Приложение

Исходные коды скриптов для построения графиков можно найти на моем github: https://github.com/dpfbop/bioinformatics