課題1

課題14-1-1

- refsnp_chrX.zip をダウンロードして展開すると、refsnp_chrXフォルダの下に、データベースファイル群 refsnp_chrX-10000.sqlite3, refsnp_chrX-100000.sqlite3, refsnp_chrX-1000000.sqlite3 , refsnp_chrX-5000000.sqlite3 , refsnp_chrX-10000000.sqlite3 が生成する。これらのデータベースのmutationsテーブルには、refsnp_chrY.sqlite3 の mutationsテーブルと同じ形式で、ヒトのX染色体において知られている突然変異のデータが格納されている。なお、それぞれのテーブルに格納されているデータ数は、10000, 100000, 1000000, 5000000, 10000000 である。

課題14-1-1

- それぞれのデータベースを以下のコマンドで検索する。

  SELECT count(*) FROM mutations

  WHERE pos BETWEEN 3000000 and 6000000;

- 適切なインデックスを作成する前と後の検索の実行時間を、DB4Sを用いて計測する。テーブルのデータ数に対して、インデックスなしの実行時間とインデックスありの実行時間をプロットし、おおむね、前者が O(n)、後者が O(log n) であることを確認せよ。
  - 測定によって得られた検索およびインデックス作成の実行時間と、インデックス作成に用いたSQLスクリプト、グラフの描画を行ったノートブックファイル(*.ipynb)を提出すること