Data pipelines and the use of NoSQL databases in genomic data

Chapter 1-Introduction

Research Statement

The purpose of this research will be to create a data pipeline from the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) to a NoSQL database. RNA sequencing (RNA-Seq) data from SRA dataset SRR1272668. The overall goal for this project is to use the SRA-toolkit and Python to download this study from the NCBI website. The data will be uploaded into a SQLite3 database using Python. Additionally, this data will also be uploaded into MongoDB. If time permits, I hope to compare the SQL and NoSQL databases search capabilities.

Background

Genetic data has become one of the fastest-growing data types in the world of Big Data. Much recent focus has been on sequencing, comparing, and annotating countless species. The transcriptome, which displays which genes are active or not, can tell us a great deal about the role of genes in our physiology and phenotypes. Next-generation sequencing (NGS) has made sequencing genomes and transcriptomes much faster and cheaper than it has been in the past, leaving biologists with a great deal of data to analyze. RNA-Seq data analysis, in particular, allows us to compare RNA sequences across groups, organisms, and species. Not only does this allow scientists to discover the phenotypic expression of genes, but they can also track the phenotypic variants of a gene.

Deliverables Statement

There will be two deliverables (hopefully!) for this project, the first being the results of a query displaying the data in a SQL database, and the second being the results of a query in MongoDB. I hope to have multiple queries to acknowledge the existence and use of my databases. If time permits, I would also like to create either a visual or written guide to assist me in the process in the future.

Chapter 2 – Technical Components

The major technologies used in this project include Python, SQLite3, and MongoDB. The SRA-toolkit, BioPython, and NumPy will also be used. The data will originate on the NCBI website, and upon completion, will reside in (hopefully) a SQL and a NoSQL database.

Potential to-do list:

Step 1: Use SRA-toolkit at the command-line level to download srr file of project SRR1272668. Use the toolkit to convert the project into FAST-Q format.

Step 2: Find a method to make the file parsable in Python. Use BioPython tutorial http://biopython.org/DIST/docs/tutorial/Tutorial.html#sec%3ASeqIO-index when necessary.

Step 3: Use Python to download the file into a SQLite3 database. If time permits, investigate on how to query/add/change this database.

Step 4: Upload the data to MongoDB. I am uncertain if I want to do this through Python. I am also not completely set on MongoDB; but I suspect this may be one of the easier NoSQL databases to attempt this on.

Step 5: (If time permits) Create either a visual or written explanation of how this process works. Also, include what methods could help in the future.

**Special note: This is a data file from the lab I will be joining this fall. My real goal with this project is to understand how to read in and work with the file. For this reason, there is a chance the focus of my research may shift. I have several books and tutorials to reference, and hope to figure out the ways these files are traditionally worked with, and how I could improve the data pipeline.