Biol 381 – Bioinformatics Lab
Lab 3: Database Searching with Biopython

Introduction

Much of research involves finding out what is already known about a topic you'd like to study. This means that it's important to develop skills that allow you to get the information you need, as well as digest that information.

Searching NCBI

Primary and secondary sources of information

NCBI contains several different databases of biological literature. PubMed is a database that contains more than 20 million citations. PubMed includes peer-reviewed reports of original research ("primary literature") and expert summaries of a topic called review articles (an example of "secondary literature"). Other sources of secondary literature found at NCBI include the Gene, Online Mendelian Inheritance in Man (OMIM) and Bookshelf databases.

Classify the effectiveness of a query

A search term's effectiveness can be determined carefully analyzing the results of the query. The records returned by your the query are all "positive" results, but they may or not all be relevant. The search results which are relevant to your search query are "true positives". The research results which are not relevant to your search query are "false positives". What about the records which were not returned by your query? These records which are not returned are "negative" results. It is also possible that your query failed to return records that are relevant, these are "false negatives". Finally, all the other articles in the database that didn't come up in your search and aren't relevant, are *true negatives*. This framework will be used throughout the course, so you should begin to familiarize yourself with the basics. For a more in-depth explanation of the concept, you can also refer to Figure 2.13 on p. 41 in your textbook.

Part 1: Searching NCBI's pubmed using Biopython [class exercise]

We'll use Biopython's Entrez library to explore NCBI's Pubmed and Gene databases. NCBI's Gene database contains a summary of information about genes including chromosomal locations, pathways, and phenotypes. NCBI's PubMed is a database containing citations and abstracts for millions of peer reviewed biology and biomedical journal articles. We'll use Biopython to write a short script to search Pubmed and and Gene for information related to cystic fibrosis.

Step 1:

Open the iPython notebook Entrez_pubmed.ipynb and write in your name.

Step 2:

Make modifications to the code and add comments to the notebook as we go through the exercise.

We’ll go through the first exercise as a class, then you’ll work with your group to complete the remaining exercises.

Due next lab period:
An iPython notebook containing your commented code.