Project Proposal

Protein Sequences to Curated Drugs

I would like to have a program where the user enters either of 2 types of entries. The first is a protein sequence let’s say from a mass spec experiment and the scientist was to see if there are already existing drugs to start the Computer Aided Drug Design (CADD) process. This program does actual perform the CADD, but potential hits. The second entry type is a Uniprot protein ID and a start and stop location. Let’s say a different scientist determines a mutation of a set of cancer patients and the mutations occur in a particular region/sequence/domain and they wanted to see if there are curated drugs for this sequence, they could also enter that information.

This is really just experimental and I will outline the simple requirements but there would need to be much more work to define more requirements such as biophysical properties of the drugs, adverse drug side effects, possible model organisms.

Functionality Workflow:

1. First sequence is gained from either entry (harder for entering UniProt as need to make call) and incorporated into fastq file. This will be done by a python cgi script.
2. Then the cgi script will run a blastp on the Swissprot/Uniprot. We have blastp installed on server. I will need to install on my Mac for testing.
3. The results returned from the blastp will be in number 6=tabular format (<https://www.biostars.org/p/88944/> plus the “staxid”) which the cgi script creates the table for with datatables. I hope that I can make it data selectable (<https://jsfiddle.net/annoyingmouse/yxLrLr8o/>) with datatables as well as filter the organism (<https://datatables.net/extensions/fixedheader/examples/options/columnFiltering.html>)
4. Then from the UniProt IDs selected, obtain all the ChEMBL drug IDs (see python package below) as well DrugBank(DB) IDs. This can be done in multiple ways.
   1. Create list of ChEMBL IDs as key for each UniProt ID by querying UniProt directly and then follow package bwlow for ChEMBL IDs. Then use chembl-data-web-services to get the sdf file under <https://chembl.gitbook.io/chembl-interface-documentation/web-services/chembl-data-web-services>

The imports and packages for this are most from <https://github.com/chembl/chembl_webresource_client>

import math

from pathlib import Path

from zipfile import ZipFile

from tempfile import TemporaryDirectory

import numpy as np

import pandas as pd

from rdkit.Chem import PandasTools

from chembl\_webresource\_client.new\_client import new\_client

import json

So we just need rdkit installed (I think can only be done with conda <https://www.rdkit.org/docs/Install.html>) as well as chembl\_webresource\_client (https://github.com/chembl/chembl\_webresource\_client).

* 1. Still with chembl\_webresource\_client but by targets such as with:

targets\_api = new\_client.target

uniprot\_id = "P00533"

targets = targets\_api.get(target\_components\_\_accession=uniprot\_id).only(

"target\_chembl\_id", "organism", "pref\_name", "target\_type"

)

Then get the ChEMBL ids that way and then with each of those follow the processes above under “[Get a single molecule by ChEMBL ID](https://chembl.gitbook.io/chembl-interface-documentation/web-services/chembl-data-web-services)”

* 1. If I have time as well can an API key from DrugBank, either get the DB ids from Uniprot such as under Chemistry databases, ex: <https://www.uniprot.org/uniprot/P00533> (many drugs here)

Then find way to get their sdf molecule files. Or again using DrugBank API search with as example (does not work without API key), with:

curl -L 'https://api.drugbank.com/discovery/v1/polypeptides/P49189'

-H 'Authorization: myapikey

* 1. Lastly if going the UniProt way to get the DB ids we can simple get the fully loaded sdf files from the drugbank url.

[https://go.drugbank.com/structures/small\_molecule\_drugs/<DB\_ID>.sdf](https://go.drugbank.com/structures/small_molecule_drugs/%3cDB_ID%3e.sdf)

<https://go.drugbank.com/structures/small_molecule_drugs/DB07662.sdf>

The database aspect would be to include a simple four table database. The entry database which will include storing the entries, so either sequences or Uniprot ID and range, and then it will have the blastp results (Also uniport IDs) in another table (limiting to 500 per entry) and from there link to another two tables for ChEMBL and DrugBank.

blastp -out P0053.xml -outfmt 5 -query P0053.fasta -db swissprot -evalue 0.001 -remote -query\_loc 200-500