Brian Wiley

Final Project

AS.410.712.81.SU21 Advanced Practical Computer Concepts for Bioinformatics

**Protein blast to drug leads**

For this project I decided to choose a topic that would relate to my future endeavors in pursuing a PhD in computational chemistry or computer aided drug design (CADD) starting next year. Proteins are the “workhorses” of all multi- and single cellular organisms. They encompass and influence all biological processes. As such “the vast majority of drugs used in medicine are targeted to proteins” (Patrick 2013). In short, this web application allows for the user to insert a peptide either obtained from some biological assay with instruments such as a protein sequencer or a mass spectrometer as well as already existing curated proteins from SwissProt to the NCBI blastp program using the SwissProt database. From the hits the user can select accession IDs and pull all the associated DrugBank IDs.

A key aspect in the drug design process is finding drug leads for a target in which can be slightly altered many times by adding various substituents and functional groups (Figure 1). This allows for testing compounds in assays with the targets to improve the absorption, distribution, metabolism, and excretion (ADME), the toxicity levels including increasing the target specificity, or the efficacy by increasing the molecule affinity for the target. Another interesting part of drug design that has resurfaced (pun to be intended soon) is not building and creating new drug compounds but discovering new targets for existing drugs. I once heard talk on cancer therapy, I forget if it was this one specifically, [New treatment horizon : Chimeric Antigen Receptor CAR T Cell Therapy](https://www.youtube.com/watch?v=gmjLSeAXAmw), but the presenter mentioned pharmaceutical research was proving new targets for “rediscovering” the use of aspirin (acetylsalicylic acid), a drug that is over 100 years old.

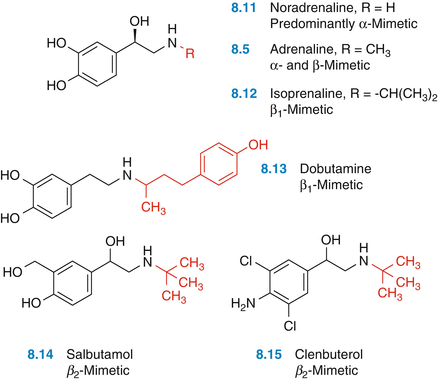


Figure 1. Adding substituents to norepinephrine to create new drug molecules that may be more specific to certain alpha and beta adrenergic receptors. *Figure taken from Optimization of Lead Structures* by *Gerhard Klebe*.

Just as compounds can be used as building blocks, I implemented an application that can also be built upon. **Specifically, this programs only works off of protein sequences, blasting against a single database, SwissProt, and only pulls from drugs from one drug database, DrugBank as drug leads**. There can be building blocks that include other macromolecules (nucleic acids, lipids, and carbohydrates), blasting more protein databases, other molecule type blast databases, or implementing non-blast searching such as structure homology, motif searching, fold recognition (RNA or proteins), or comparative modelling, as well as including many other drug databases such as ZINC, PubChem, and ChEMBL/ChEBI.

I hope this program will be able to influence other ideas as well for searching proteins which are functionally and structurally similar to existing proteins in where there are successful and efficacious therapies. Some of the challenges I endured in my development included are as follows. Determining the scope was a challenge. Despite being able to summarize the program in one sentence (above in bold), I initially wanted to search many chemical databases such as those indicated above as well and also included some functionality of the REACTOME Pathway Browser which includes Chemical Compounds in their pathways, for example these [12 compounds](https://reactome.org/PathwayBrowser/#/R-HSA-1227986&DTAB=MT) for this EGFR pathway. As I only had a month time, mostly just limited to weekends, I had to narrow down the scope. Another challenge comes from the fact that from the DrugBank database, peptide drugs such as **Exenatide**, and monoclonal AntiBody (mAB) drugs do not even have links to PDB files. Despite including the fasta sequences, the molecular geometry and publication of PDB files are limited for these proprietary type drugs. An example of a minor challenge sort of sums up typically challenges you see with implementing JavaScript libraries in your applications. For the checkboxes library I used Gyrocode’s jquery-datatables-checkboxes, where one limitation is that the select-all can only select all items visible on the current page. So, if you used DataTable pagination you were limited to only selecting all items on the current page. To solve this challenge, I implemented vertical scrolling with all hits on the first/only page.

Some highlights I would like to summarize are as follows. I like the way to program looks. It is sleek and simple with really only 2 pages for the user. I like simple. For instance, for the RCSB PDP interfaces I really only use 4 pages: the search, the selection, the entry, and the annotations tab. I like that it builds off another application to do an additional task. After it runs NCBI Blastp is gets SDF files of associated drugs with those hits. I also like that it can save the results to a tar file. This is pretty convenient for storage. I got this idea from the zip folder that is returned for multiple entries for searching karyotype strings with CytoGPS (created by people at WashU where I work, <http://cytogps.org/karyotype_parser>). I included a document Final\_Challenges\_Highlights.docx that summarizes these challenges and highlight with images.

Things to improve upon in my application are the summary of the results. This seems kind of scattered. I included a tar file that has a summary JSON as well as folders for each accession ID and their respective DrugBank SDF files and a DrugBank fetch summary JSON file in each accessionID folder. The results are also submitted to a MySQL table peptide\_drug\_searches based off of primary key of Query\_accession\_version, Subject\_accession\_version, query\_start residue, query\_end residue, subject\_start residue, subject\_end residue and include a JSON string for the DrugBank drug IDs and names. It is hard to come up with a summary or story of the results for the user submission but will more or less be best used by a developer and not an IT manager or high-ranking executive of a drug lead development team.

I learned a lot about the difficulties of web applications. For me the hardest things were first off how Apache handles the server and client side of web applications. This is different for different servers and computer operating system types. JavaScript is hard because it’s hard to troubleshoot the code as they are stored as one line or a few lines and so there is will be an error like jQuery.js:2 for line 2 but line 2 is thousands of characters and many functions long. Permission for CGI scripts are complex. It is helpful to include hard coded form entries for debugging from the Python interpreter side. Despite JavaScript being hard, CSS is harder. I could spend over an hour or more trying to change color or style of one complex attribute or tag. It is even harder when you have a base like my slate bootswatch and you need to override it. Here you must use complex selectors which, despite that they work and are helpful, simple documentation out there just doesn’t cut it. All in all, I am happy with my final project of my final class of my Masters in Bioinformatics at Johns Hopkins University. Thanks, and Godspeed!

References:

1. Patrick, G.L. An Introduction to Medicinal Chemistry. 5th Edition. 2013
2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. [PubMed](https://www.ncbi.nlm.nih.gov/pubmed/2231712?dopt=Citation)
3. The UniProt Consortium  
   **UniProt: the universal protein knowledgebase in 2021**  
   [Nucleic Acids Res. 49:D1 (2021)](https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa1100/6006196)
4. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006 Jan 1;34 (Database issue):D668-72. 16381955.