DATA SCIENCE FUNDAMENTALS

RANA MORRIS // *Data Science Fundamentals Capstone Project*

PROTOTYPE DRUG DISCOVERY PIPELINE PROJECT: HUMAN GROWTH FACTOR RECEPTORS AS CHEMOTHERAPY TARGETS

QUESTION // PROBLEM

NLM Resources support the full span of lifecycle of medications development: from compound discovery to biological activity assessment to gene targeting to disease/disorder therapeutic selection.

Project Goal: Create a prototype pipeline for the discovery of novel gene-targeting drug candidates.

Potential Impact: Drug candidate recommendations could be shared to assist researchers with an evidence-based starting point for new drug discovery.

DATA & MODELS

PubChem Compound & BioAssay Data, retrieved from new APIs, focusing on:

- 'geneid' (NCBI Gene IDentifier - for the gene product I am targeting)
- 'cid' (PubChem Compound Identifier - for the chemical compounds I am examining)

VARIABLES

Assessment variables:

- 'acname' (the activity assessed in the BioAssay data for the 'geneid's)
- 'acvalue' (the measured activity value of 'cid's from the BioAssay experiments)

Filtering variables:

- 'collection' ('cid's identified as drugs by drugbank or dailymed were used as positive controls and eventually removed from the lead compound list)
- 'geneid' ('cid's identified with low 'acvalue's for non-gene target 'geneid's were removed from the drug candidate list)

CONCLUSION // NEXT STEPS

This pipeline identified 5 drug candidates for each of 14 human growth factor receptors based on good inhibitory or binding effectiveness and limited cross-reactivity with other gene products.

These drug candidates are recommended for the next stage of drug discovery, experimental validation:

- *in vitro* with purified target
- *in vitro* with a cultured cancer cell line
- *in situ* with a model organism
- and ultimately in clinical trials in humans.

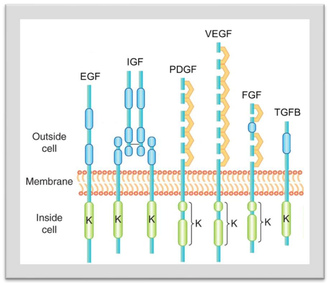
This code can be cleaned up, fully automated, and run to find potential drug candidates for other 'geneid's based on constantly updating data from PubChem.

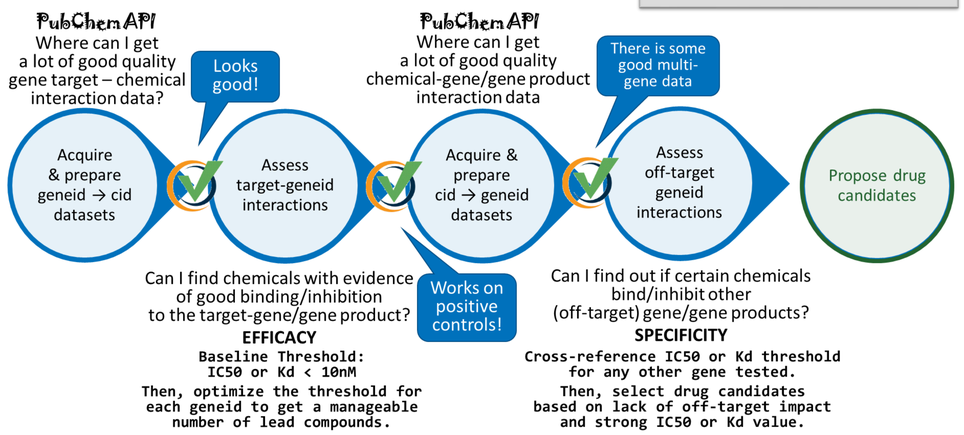
Project Plan

This project aims to identify good drug candidates for each of 14 human growth factor receptors.

This gene family is often found to be hyperactive in many cancers. Thus, these are good drug targets for the design or discovery of inhibitory compounds.

Several of these gene products are already targeted by FDA-approved drugs, which served as a positive control for this pipeline's relevancy and effectiveness.





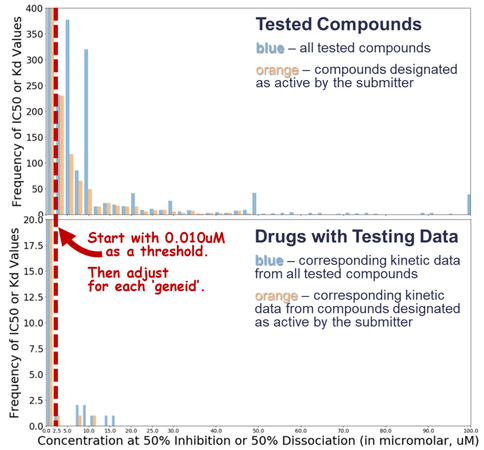
Which measure and threshold should I use to select lead compounds?

Bioactivity data measuring each 'cid's impact on the target 'geneid' was most often determined as an IC50 or Kd value ('acname').

acname	Activity	EC50	IC50	INH	Kd	Ki	Km	Selectivity
geneid								
1956	0	68	622	0	227	2	0	0
2260	0	0	229	0	20	0	0	0
2261	0	0	166	0	40	0	0	0
2263	1	0	1208	0	137	11	0	0
2264	0	0	22	0	20	0	0	0
2321	0	0	300	0	21	1	0	0
2324	0	0	13	0	21	0	0	0
3480	0	0	1356	0	20	0	5	0
3482	0	0	12	0	0	0	0	0
3791	0	15	2146	4	26	5	0	3
5156	0	0	61	0	21	0	0	0
5159	0	0	139	0	20	0	0	0
7046	0	0	181	1	20	0	0	0
7048	0	0	77	0	123	0	0	0

A 'cid's potential impact on a 'geneid' was initially assessed with IC50 or Kd bioactivity threshold values of 0.010 micromolar ('acvalue').

Then, 'geneid'-specific adjustments of the threshold 'acvalue' were used to assemble a list of 8-16 lead compounds per 'geneid'.



Which lead compounds have limited off-target effects and good activity?

BioAssay data for lead compounds against *all* 'geneid's was retrieved and filtered based on each target 'geneid's specific threshold.
An example of lead compounds targeting 'geneid' 5159 is shown below:

The number of assays indicating kinetics better than the 0.1uM threshold is:

geneid	25	27	2050	2185	2322	2872	3791	3815	4233	4486	...	4921	5156	5159	5605	6714	7010	7525	8767	112
cid																				
11569758	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	1	0	0	0	0
11719421	0	0	1	0	0	0	0	0	0	0	...	0	0	1	0	1	0	1	0	0
42642645	0	2	0	2	4	2	10	2	8	6	...	2	4	4	2	0	2	0	2	2
44444041	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	1	0	0	0	0
46831579	1	0	0	0	0	0	0	1	0	0	...	0	0	1	0	0	0	0	0	0
118130052	0	0	0	0	0	0	1	0	0	0	...	0	0	1						
118130056	0	0	0	0	0	0	1	0	0	0	...	0	0	1						
118130067	0	0	0	0	0	0	1	0	0	0	...	0	0	1						
118130077	0	0	0	0	0	0	1	0	0	0	...	0	0	1						
118130111	0	0	0	0	0	0	1	0	0	0	...	0	0	1						
118699724	0	0	0	0	0	0	2	0	0	0	...	0	0	1						

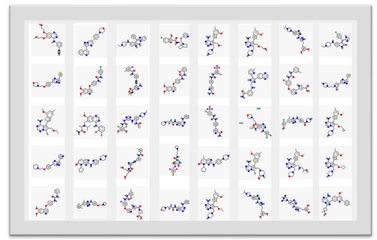
Drug candidates were selected based on limited cross-reactivity (left) and high strength of activity ('acvalue').

	cid	aid	activity	acname	acvalue	geneid
68	44444041	1342658	Active	IC50	0.0012	5159
70	11569758	1342658	Active	IC50	0.0028	5159
3	11719421	1342658	Active	IC50	0.0059	5159
77	118130052	1343857	Active	IC50	0.0060	5159
79	118130067	1343857	Active	IC50	0.0060	5159
71	118699724	1343779	Active	IC50	0.0070	5159
75	118130111	1343857	Active	IC50	0.0070	5159
83	118130077	1343857	Active	IC50	0.0080	5159

Pipeline Results

Five potentially good drug candidates were identified for each of the 14 human growth factor receptors.

These were chosen based on publically-available PubChem BioAssay data showing strong inhibitory or binding potential and for limited interaction with other genes/gene products.



96,105,154 Chemicals in the PubChem Compound database
← 3,469,202 with BioAssay data

human gene symbols & names	geneid	number of 'cids' tested	number of tested 'cids' with kinetic data	number of lead compounds	IC50 or Kd threshold (uM)	number of drug candidates	IC50/Kd range of drug candidates (nM)	number of off-target genes tested
EGFR - epidermal growth factor receptor	1956	23,503	919	16	0.00300	5	1.0 - 2.0	49
FGFR1 - fibroblast growth factor receptor 1	2260	7,557	249	16	0.00150	5	0.50 - 0.9	7
FGFR2 - fibroblast growth factor receptor 2	2263	2,960	1,359	16	0.00045	5	0.10 - 2.0	11
FGFR3 - fibroblast growth factor receptor 3	2261	5,890	206	15	0.00300	5	0.20 - 2.5	2
FGFR4 - fibroblast growth factor receptor 4	2264	2,560	42	11	2.0	5	0.20 - 1,500	177
IGF1R - insulin like growth factor 1 receptor	3480	8,334	1,381	8	0.0100	5	3.0 - 8.0	6
IGF2R - insulin like growth factor 2 receptor	3482	20	12	7	6.0	5	850 - 4,760	3
PDGFRA - platelet derived growth factor receptor alpha	5156	5,314	82	14	0.0100	5	2.0 - 5.0	383
PDGFRB - platelet derived growth factor receptor beta	5159	4,595	159	12	0.0100	5	0.20 - 6.0	386
TGFBR1 - transforming growth factor beta receptor 1	7046	2,533	202	10	0.0200	5	6.0 – 130	283
TGFBR2 - transforming growth factor beta receptor 2	7048	213	200	14	0.400	5	26 – 230	422
(VEGFR1) FLT1 - fms related tyrosine kinase 1	2321	5,933	322	14	0.00600	5	2.0 - 4.0	7
(VEGFR2) KDR - kinase insert domain receptor	3791	17,740	2,199	16	0.00045	5	0.06 - 2.0	17
(VEGFR3) FLT4 - fms related tyrosine kinase 4	2324	3,088	34	8	0.1600	5	18 - 29	176