Predicting cancer resistance mutations with protein-ligand graph networks

Robert Heeter[1,2], **Stephanie Schmidt**[1], **Bissan Al-Lazikani**[1]

[1]*Department of Genomic Medicine, MD Anderson Cancer Center,* [2]*Department of Bioengineering, Rice University*

rch5@rice.edu • *Submitted August 8, 2022*

Motivation and goals

Drug resistance

- Generic cytotoxic treatments for cancer can include serious side effects; targeted therapies are a promising alternative, but can fail due to acquired drug resistance
  - >3000 unique mutations in >400 cancer genes have been recorded in the OncoKB database, including those that confer resistance[1]
- Resistance can be acquired through many different signaling and epigenetic pathways, though drug target mutations can be a direct cause for resistance[2]
  - EGFR becomes resistant to inhibitors gefitinib and erlotinib within ~1 year due to amino acid mutations[2]
- The ability to accurately predict resistance-conferring mutations would vastly improve treatment
  - Combinatorial or sequential therapies to circumvent resistance[2]
  - Simultaneous development of multiple drugs to handle resistance (see **Figure 1**)

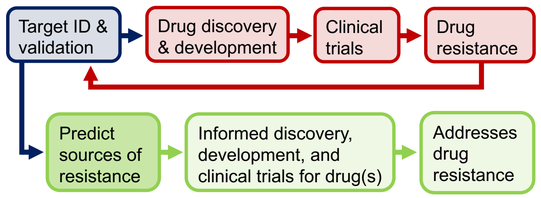


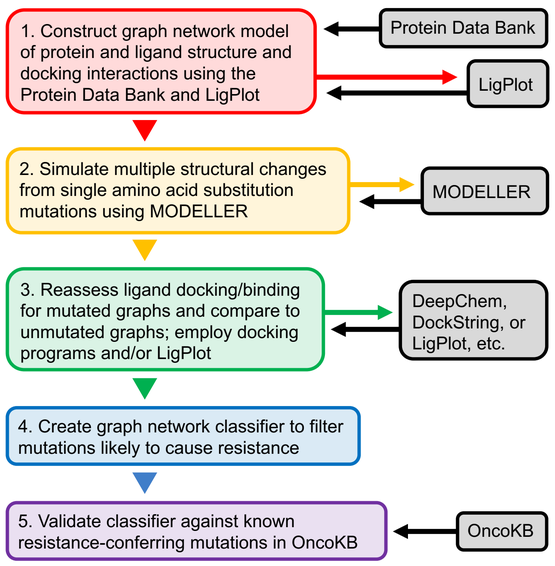
Figure 1. *Simultaneous development of multiple drugs to handle resistance following resistance prediction (shown in green); red indicates traditional pathway.*

Graph networks

- Graph networks are a promising new way to model and analyze protein-ligand structures
  - Atoms as nodes; bonds and interactions as edges

Program pipeline

- Program developed in stages using JupyterLab and Visual Studio Code
- Designed to be modular to allow flexibility in usage



1. Protein-ligand graph network construction

- First, the PDB protein structure must be parsed and bonds/interactions must be added to produce the network
  - The program adds polypeptide covalent bonds (primary structure) (see **Figure 4**)
  - LigPlot is responsible for ligand covalent bonds, hydrogen bonds, and hydrophobic interactions (see **Figure 4**)

Protein Data Bank (PDB)

- Contains atomic information, such as coordinates, for each atom in the protein
- Does not explicitly include amino acid bond information
- PDB data is converted to an atom list data structure (see **Figure 2**)
- Atom list used to produce connection (adjacency) matrix using amino acid chemical structures
- Connection matrix converted to tuple edge list (bond list)

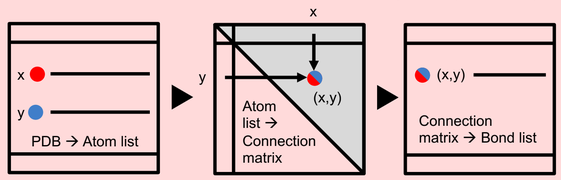


Figure 2. *PDB data manipulation to produce bond list.*

LigPlot

- Made up of 3 separate programs to identify and visualize hydrogen bonds and hydrophobic interactions between a protein chain and ligand (see **Figure 3**)
- Identifies hydrogen donors and acceptors on the amino acid chain, and uses geometry (distances and angles) to determine where bonds/interactions exist
  - Hydrogen bonds between 2.70 and 3.35 Angstroms
  - Hydrophobic interactions between 2.90 and 3.90 Angstroms
- Also uses the PDB Het Group Dictionary to assign bond orders to ligand bonds

Figure 3. *Example LigPlot ouput for the EGFR kinase domain with IRE gefitinib.*

Hydrophobic interactions are shown as "eyelashes" and hydrogen bonds are shown as dashed lines.

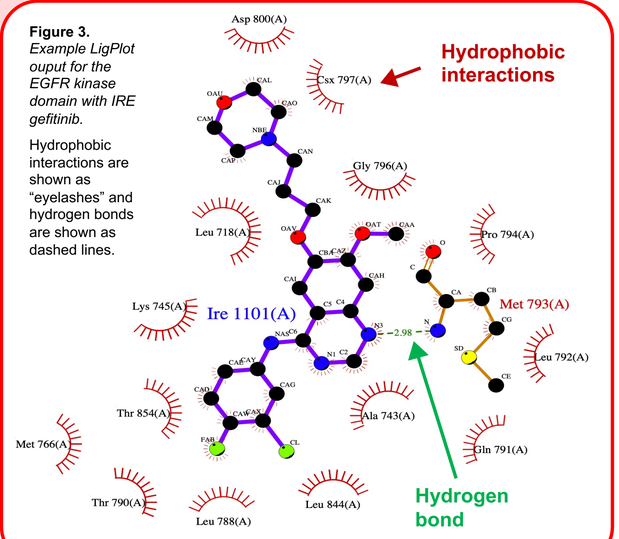


Figure 4. *3D plot of the EGFR kinase domain with IRE gefitinib using Python's Matplotlib library from graph network data structures.*

While not directly useful in the program pipeline, this 3D plot serves as a checkpoint for validating the graph network. The annotations denote the bond coloring scheme.

Note the single hydrogen bond between a nitrogen on gefitinib and Met #793 on EGFR (which matches experimental observation), along with numerous hydrophobic interactions, all determined using LigPlot.

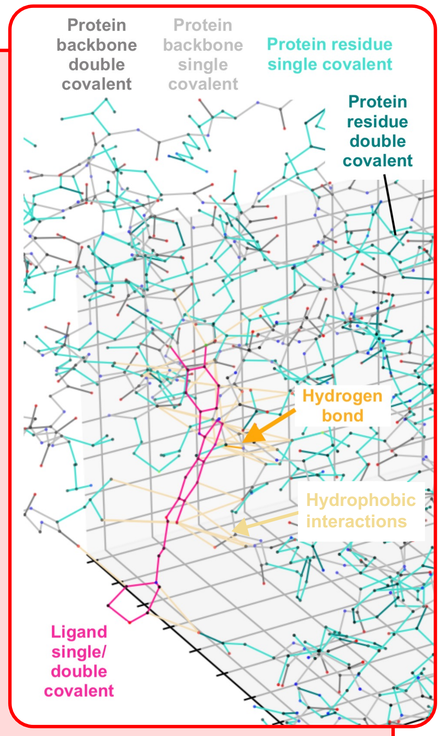
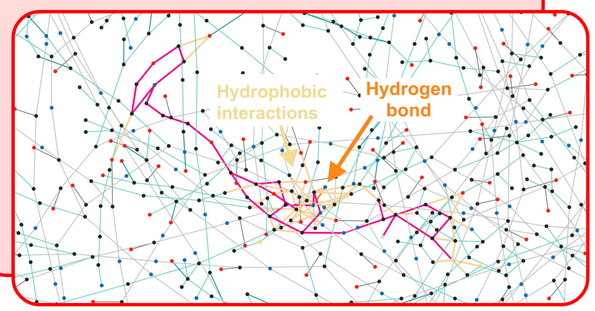


Figure 5. *Cytoscape network model of the EGFR kinase domain with IRE gefitinib.*

The bond coloring scheme matches that in *Figure 3*.



2. Mutation simulation

- MODELLER program (run through Python) used to simulate single amino acid substitution mutations around the docking site (see **Figure 6**)
  - Uses alignment of a mutated target sequence to a template sequence to modify the residues in the template PDB
  - Then, optimizes spatial restraints (energies)
  - Iterates, each time returning a modified PDB file and energy profile, used to gauge model accuracy
- Ligands can either be included as a spatial restraint (optimization around the ligand) or removed (mutant protein relaxes in the absence of the ligand)
  - Both cases are worth further study
  - For testing, the ligand IRE gefitinib was included in the EGFR kinase domain (see **Figure 6**)

3. Binding reassessment

- Determine if new mutant PDBs still permit ligand docking or if mutants are resistant, potentially using LigPlot to recalculate binding interactions
- Other docking software may be needed to optimize ligand orientation (i.e. DeepChem or DockString)

4. Network classifier

- Design classifier to distinguish between effective and resistant protein-ligand graph networks using MODELLER mutants as a training set
- Consider using a graph neural network approach

5. Classifier validation

- Use OncoKB database of known resistance mutants to test classifier

Summary and next steps

- Robust construction of a protein-ligand graph network given an input PDB file using LigPlot
- Working mutation simulation (including ligand) using MODELLER, though may want to remove ligand
- More work needs to be done to reassess binding (perhaps using docking software) and classify the new mutated models as drug resistant or not
- Need to validate the entire pipeline, especially given the use of many dependencies

References

1. Chakravarty, D., Gao, J., *et al.* (2017). OncoKB: a precision oncology knowledge base. *JCO precision oncology*, *1*, 1-16. doi.org/10.1200/PO.17.00011
2. Martínez-Jiménez, F., Overington, J. P., Al-Lazikani, B., & Marti-Renom, M. A. (2017). Rational design of non-resistant targeted cancer therapies. *Scientific reports*, *7*, 46632. doi.org/10.1038/srep46632

Acknowledgements

A special thanks to my project leader Dr. Stephanie Schmidt and lab P.I. Professor Bissan Al-Lazikani for their wisdom and support during my summer experience, as well as Dr. Phillip Gingrich for his input and feedback.

```
>P1;4wkq_template
structure:4wkq_mod.pdb   :FIRST:@:  END: :::     :
AMGEAPNQALLRILKETEFKKIKVLGS/--GTVYKGLWIPEGEKVKIPVAIKE/----SPKANKEILDEAYVMAS
VDNPHVCRLLGICLTSTVQLITQLMPFG/LLDYVREHKDNIGSQYLLNWCVQIAKGMNYLEDRRLVHRDLAARNV
LVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWMALESILHRIYTHQSDVWSYGVTVWELMTFGSKPYDGI
PASEISSILEKGERLPQPPICTIDVYMIMVKCWMIDADSRPKFRELIIEFSKMARDPQRYLVIQGD/--------
------------MDDVVDADEYLIPQ/.*

>P1;4wkq_target
sequence :4wkq_target:     : :     : :::     :
AMGEAPNQALLRILKETEFKKIKVLGS/--GTVYKGLWIPEGEKVKIPVAIKE/----SPKANKEILDEAYVMAS
VDNPHVCRLLGICLTSTVQLITQLWPFG/LLDYVREHKDNIGSQYLLNWCVQIAKGMNYLEDRRLVHRDLAARNV
LVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWMALESILHRIYTHQSDVWSYGVTVWELMTFGSKPYDGI
PASEISSILEKGERLPQPPICTIDVYMIMVKCWMIDADSRPKFRELIIEFSKMARDPQRYLVIQGD/--------
------------MDDVVDADEYLIPQ/.*
```

Figure 6. *Template-target alignment sequence of the EGFR kinase domain with IRE gefitinib for MODELLER.* A mutation from Met (M) #793 to Trp (W) (highlighted) was simulated in this case. The "/." at the end of the sequence indicates a ligand.