Python Ligand Fit Tool

1. Background

Virtual screening is a critical tool in drug discovery, used to predict how potential ligands bind to target proteins. In this study, we focused on Riboflavin Synthase (RibF), virtually docking numerous ligands using AutoDock Vina. However, Vina has a known limitation: it tends to favor larger molecules, often assigning them high docking scores even when they don't fit well within the binding pocket. These mis-docked ligands may stick out or fail to bind snugly, leading to biased results.

To address this issue, we developed a tool to flag these potentially mis-docked ligands. By identifying ligands that receive artificially high scores despite poor spatial fit, this tool improves the reliability of virtual screening outcomes, helping to focus experimental efforts on more promising candidates.

2. Product

2.1 Determining Fit Data

A Python program was created to help determine ligand fits into a protein taking in a PDB file for the protein and either a PDB or a PDBQT file for the ligand(s). Protein atoms were placed in a KD Tree data structure in order maintain *O(log(n))* time complexity for finding the nearest protein atom. Iterating through each ligand atom, the nearest protein atom was found by using the Euclidean distance formula between the ligand and protein atom. These were stored in a Python Pandas Series and as such, could contain additional calculated metadata. Namely, the 5 data points that were taken and stored were the median, minimum, maximum, Q1, and Q3 values. These parameters were chosen to represent the entire ligand's fit into the protein. These series were then combined into a Pandas DataFrame for ligand comparison.

2.2 Known Ligand-Protein Comparisons

Using the first 20 protein-ligand complexes in the DUD-E (Database of Useful Decoys-Enhanced)[1] set of known pairs, the same fit data DataFrame was calculated for these pairs. The DataFrame is shown below.

Table 1: DUD-E Fit Data

PDB File	Mean (Å)	Median (Å)	Max (Å)	Min (Å)	Q1 (Å)	Q3 (Å)
3eml	3.573233	3.465908	4.88329	2.997396	3.258394	3.720942
3bkl	3.57948	3.591736	4.44878	2.34495	3.466395	3.91357
1e66	3.525178	3.554598	3.82616	2.914779	3.426745	3.634649
2e1w	3.585742	3.671133	4.20148	2.733374	3.472957	3.834322
2oi0	3.619092	3.598886	4.64099	2.871849	3.271804	3.883665
3cqw	3.387871	3.534801	3.97415	2.55632	3.167302	3.66808
2am9	3.768525	3.820393	4.64042	2.662012	3.457478	4.181073
1bcd	3.308862	3.183841	4.01148	2.742904	3.058207	3.5583
2cnk	3.046764	3.263759	3.91834	1.309845	2.768301	3.496224
3ny8	3.405322	3.465541	3.85904	2.651094	3.173171	3.687336
2hv5	3.442355	3.428787	4.25734	2.68562	3.26379	3.674317
1h00	3.511969	3.484266	5.80968	1.785379	3.199728	3.764885
2hzi	3.607852	3.545453	4.16780	2.921677	3.423309	3.836971
2vt4	3.477357	3.564625	4.13370	2.565209	3.277269	3.703171
3d0e	3.525986	3.510288	4.10712	2.641468	3.37445	3.774441
1s3b	3.580241	3.579268	4.08284	3.055219	3.32699	3.78849
3l5d	3.857426	3.733828	6.12861	2.603711	3.39863	4.227282
3d4q	3.567129	3.644103	4.34916	2.084746	3.340156	3.790205

From these, a baseline of these data parameters was determined. The 2 main parameters for which baselines were established included the mean and maximum values for the ligand. From the DUD-E data, we determined that a mean value of greater than 3.8 Å or a maximum value of greater than 5 Å would flag a poorly docked ligand. These determinations were confirmed upon visual analysis of docked ligand through Chimera, a tool to visualize PDB files.

3. Future Direction

This tool can evaluate the fit data for any protein-ligand complex by analyzing its PDB file, making it applicable across a wide range of docking scenarios. The current parameter cutoffs can be applied to docked ligands, such as those in RibF, to identify potentially successful docking interactions. These predictions can then be validated through experimental testing, confirming the accuracy of virtual screening results.

Additionally, the parameter cutoffs may be refined over time. By analyzing a larger and more diverse set of protein-ligand complexes, the tool may reveal new or optimized thresholds that provide more precise assessments of ligand docking quality. This continuous refinement will enhance the tool's accuracy and broad applicability in drug discovery efforts.

4. References

1. https://dude.docking.org/