Python Ligand Fit Tool

Anthony Shatby

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.[2] 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 extracted from the PDB file and placed into a KD Tree. By using the KD Tree data structure, *O(log(n))* time complexity was maintained for finding the nearest protein atom, where n is the number of atoms in the protein.[3] 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 17 protein-ligand complexes in the DUD-E (Database of Useful Decoys-Enhanced)[1,4] set of known pairs, the same fit data DataFrame was calculated for these pairs. The ligand protein complexes (Table 1) and generated DataFrame (Table 2) are shown below.

Table 1: DUD-E Library Protein Ligand Complexes

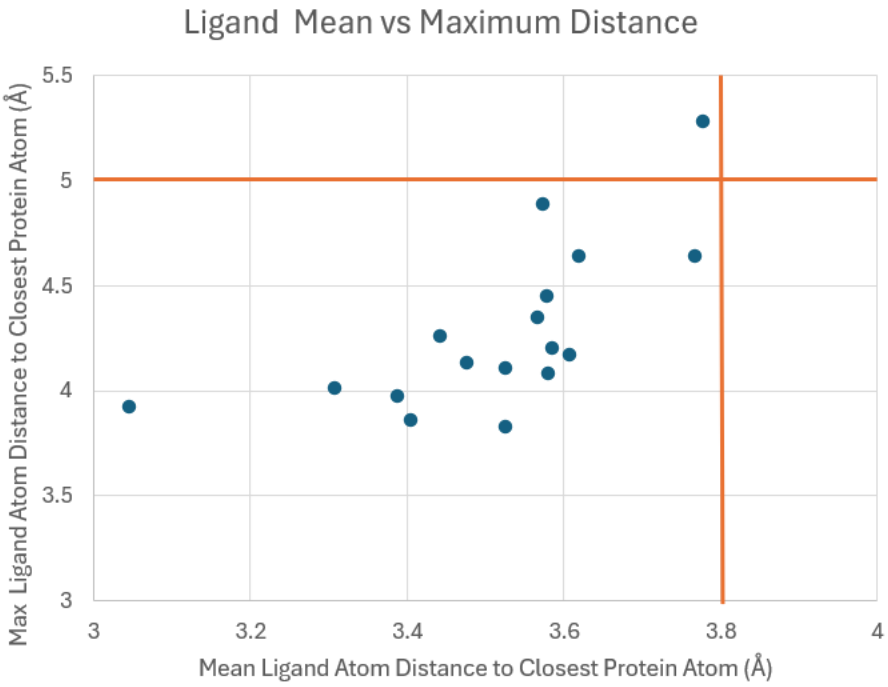
PDB File	Protein Name	Ligand PDB Code
3eml	Human Adenosine A2A receptor	ZMA
3bkl	Angiotensin-converting enzyme, somatic isoform	KAW
1e66	Acetylcholinesterase	HUX
2e1w	Adenosine deaminase	FR6
2oi0	TNF- a Converting Enzyme (TACE)	283
3cqw	RAC-alpha serine/threonine-protein kinase	CQW
2am9	Human Androgen Receptor	TES
1bcd	Carbonic anhydrase II	FMS
2cnk	Caspase-3 p17 subunit	MY2
3ny8	Human Beta2 Adrenergic Receptor	JRZ
2hv5	Human Aldose Reductase	ZST
2hzi	Proto-oncogene tyrosine-protein kinase ABL1	JIN
2vt4	Beta1 Adrenergic Receptor	P32
3d0e	RAC-beta serine/threonine-protein kinase	G93
1s3b	Amine oxidase [flavin-containing] B	RMA
3l5d	Beta-secretase 1	BDV
3d4q	B-Raf proto-oncogene serine/threonine-protein kinase	SM5

Table 2: DUD-E Ligand Fit Data

PDB File	Mean (Å)	Median (Å)	Max (Å)	Min (Å)	Q1 (Å)	Q3 (Å)
3eml	3.573	3.465	4.883	2.997	3.258	3.720
3bkl	3.579	3.591	4.448	2.344	3.466	3.913
1e66	3.525	3.554	3.826	2.914	3.426	3.634
2e1w	3.585	3.671	4.201	2.733	3.472	3.834
2oi0	3.619	3.598	4.640	2.871	3.271	3.883
3cqw	3.387	3.534	3.974	2.556	3.167	3.668
2am9	3.768	3.820	4.640	2.662	3.457	4.181
1bcd	3.308	3.183	4.011	2.742	3.058	3.558
2cnk	3.046	3.263	3.918	1.309	2.768	3.496
3ny8	3.405	3.465	3.859	2.651	3.173	3.687
2hv5	3.442	3.428	4.257	2.685	3.263	3.674
2hzi	3.607	3.545	4.167	2.921	3.423	3.836
2vt4	3.466	3.563	4.137	2.529	3.303	3.726
3d0e	3.525	3.510	4.107	2.641	3.374	3.774
1s3b	3.580	3.579	4.082	3.055	3.326	3.788
3l5d	3.777	3.757	5.276	2.573	3.358	4.257
3d4q	3.567	3.644	4.349	2.084	3.340	3.790

The data from the above table was plotted in order to compare ligands.

Figure 1: Plot of Ligand Mean vs Maximum Distance



From these, a baseline of these data parameters was determined. The two 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. The orange lines on the plot display these values relative to the calculated ligand values from the DUD-E Library. 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/

2. O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, *Journal of Computational Chemistry 31* (2010) 455-461.

3. Moore, A.W. (1998). An introductory tutorial on KD-trees.

4. Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry* 2012 55 (14), 6582-6594.