Analyzing Patterns of Computational Similarity between Kinase Ligands

Jack Ringer

University of New Mexico
8/4/2025

Background and Overview

Background

There are 8 major groups within the human kinome, which include: AGC, CAMK, CK1, CMGC, STE, TK, TKL, Other, as well as 13 atypical families [2, 5]. Aside from the "Atypical" and "Other" groups, these group classifications are generally based on sequence similarity, evolutionary conservation, and known functions.

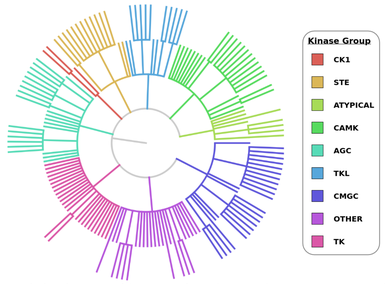


Figure 1. *Phylogenetic Tree of the human kinome showing major groups as well as families and subfamilies. Generated using ETE4 with data from ChEMBL [6].*

The similarity property principle (SPP), has been enormously influential in the realm of medicinal chemistry [4]. According to the SPP, structurally-similar compounds often exhibit similar properties (including biological activity).

This work investigates whether ligands which are active within a particular protein kinase group are more similar to one another than protein kinase ligands generally.

Why is this important?

- Relevant to drug discovery research
- If there is a relationship between protein kinase group and ligand similarity, then it may be informative to look at the ligands of related protein kinases (i.e., those belonging to the same group)

Dataset and Methodology

The set of active ligands and their relationship to specific human protein kinases was determined using data from single protein target binding assays in the ChEMBL database [6].

- Assay/ligand selection based on Pharos [3]
- Remove assays where target was a variant/mutant
- Filtered out PAINS compounds [1]
- Molecular weight of ligand must fall between [200, 900] Da

The criteria above resulted in a dataset with the folllowing properties:

Variable	Value
N. Protein Targets	423
N. Assays	73,487
N. Active Samples	38,622
N. Unique Ligands	9,995

Ligands are considered active within a protein kinase group if they were identified as an active within an assay targeting a protein belonging to the group.

After determining the set of ligands and their group relationship(s), Morgan fingerprints were computed using RDKit. Tanimoto similarity coefficients were then computed between these 2D fingerprints. Figure 2 shows the 2D structures of some ligand pairs and their similarity values.

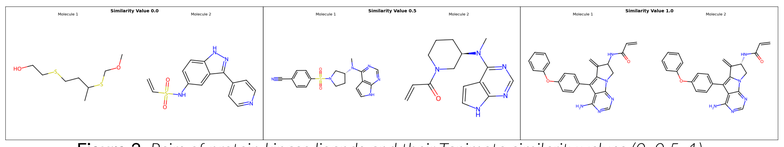


Figure 2. *Pairs of protein kinase ligands and their Tanimoto similarity values (0, 0.5, 1)*

Results

Figure 3 provides distributions for the $\binom{N}{2}$ similarity values per group (N = number of ligands), as well as the $\binom{9{,}995}{2} = 49{,}945{,}015$ similarity values calcualted for all protein kinase ligands in the dataset. Table 1 provides additional statistics for each group and results from a Mann-Whitney U test (MWUT).

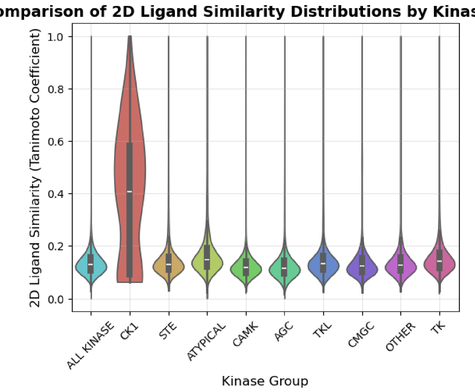


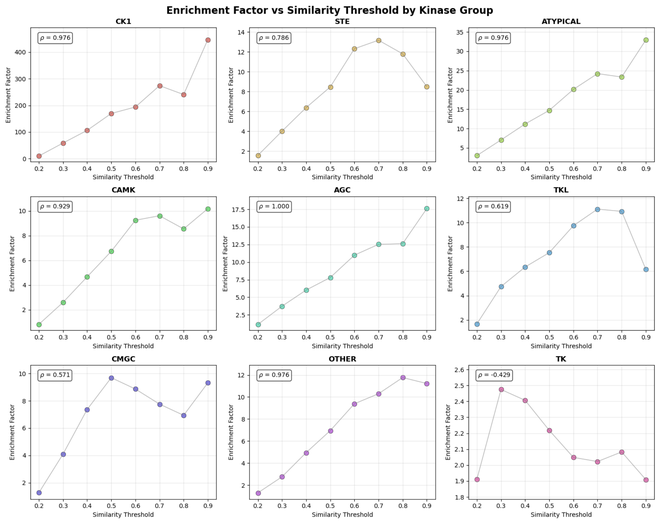
Figure 3. *Comparison of 2D structural similarity distributions by protein kinase group*

Kinase Group	N. Targets	N. Ligands	Group Median	Comparison Median	MWUT p-value
CK1	10	13	0.407	0.129	3.10×10^{-11}
STE	45	425	0.131	0.129	2.48×10^{-169}
ATYPICAL	15	357	0.149	0.129	$< 5 \times 10^{-324}$
CAMK	65	597	0.117	0.130	1.0
AGC	59	809	0.116	0.131	1.0
TKL	37	810	0.133	0.129	$< 5 \times 10^{-324}$
CMGC	58	1275	0.122	0.130	1.0
OTHER	56	727	0.127	0.129	1.0
TK	80	5347	0.140	0.121	$< 5 \times 10^{-324}$

Table 1. *Table showing comparisons of targets, ligands, and ligand similarity distributions per protein kinase group. Shown p-values are calculated (with Bonferroni correction) from a MWUT where the alternative hypothesis is that the similarity values within the group are stochastically greater than the distribution of all other similarity values.*

Figure 4 below shows the enrichment values observed for different protein kinase groups, where enrichment is defined as:

$$\frac{P(\text{two ligands active within group} \mid \text{similarity} > \text{threshold})}{P(\text{two ligands active within group})}$$



Discussion

Limitations

- Not all protein targets are equally well-studied (see Figure 5)
- Only considered Morgan fingerprints + Tanimoto coefficients when measuring "similarity" of ligands
- Methodology does not directly account for differences in assay conditions, or the fact that even single proteins can have multiple binding sites
- Data gathered from ChEMBL may not reflect global trends

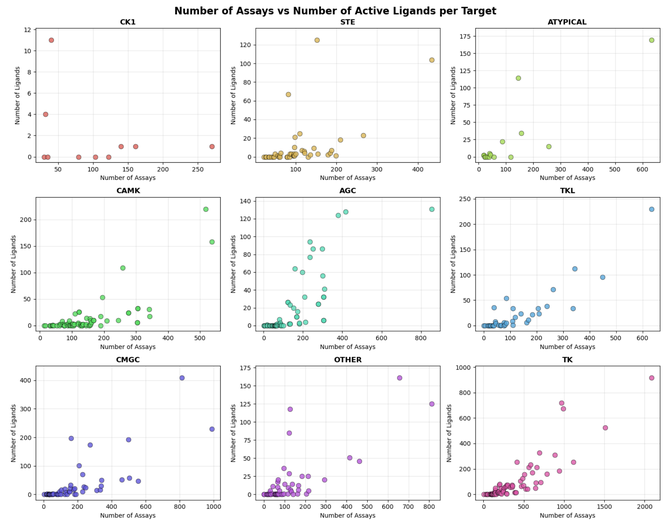


Figure 5. *Scatter plots showing the number of assays (x-axis) and ligands (y-axis) for each protein belonging to each protein kinase group.*

Conclusions

Overall, this study found no clear relationship between 2D ligand similarity and protein kinase group activity. Although significant differences were observed for the CK1 group, given the limitations outlined above further study is necessary before accepting these results.

Acknowledgements

Much thanks to my advisor Dr. Vincent Metzger for his guidance and input over the course of this project. I'd also like to thank Dr. Jeremy Yang, Dr. Cristian Bologa, and Dr. Praveen Kumar for their feedback during weekly meetings over the course of the internship. Finally, I'd like to thank the authors of ChEMBL DB [6].

References

[1] Jonathan B. Baell and Georgina A. Holloway.
New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays.
Journal of Medicinal Chemistry, 53(7):2719–2740, Apr 2010.

[2] Sameh Eid, Samo Turk, Andrea Volkamer, Friedrich Rippmann, and Simone Fulle.
Kinmap: a web-based tool for interactive navigation through human kinome data.
BMC Bioinformatics, 18(1), Jan 2017.

[3] Keith J Kelleher, Timothy K Sheils, Stephen L Mathias, Jeremy J Yang, Vincent T Metzger, Vishal B Siramshetty, Dac-Trung Nguyen, Lars Juhl Jensen, Dušica Vidović, Stephan C Schürer, Jayme Holmes, Karlie R Sharma, Ajay Pillai, Cristian G Bologa, Jeremy S Edwards, Ewy A Mathé, and Tudor I Oprea.
Pharos 2023: an integrated resource for the understudied human proteome.
Nucleic Acids Research, 51(D1), Nov 2022.

[4] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath.
Molecular similarity in medicinal chemistry.
Journal of Medicinal Chemistry, 57(8):3186–3204, Nov 2013.

[5] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam.
The protein kinase complement of the human genome.
Science, 298(5600):1912–1934, 2002.

[6] Barbara Zdrazil, Eloy Félix, Fiona Hunter, Emma Manners, James Blackshaw, Sybilla Corbett, Marleen De Veij, Harris Ioannidis, David Méndez, Juan F Mosquera, Maria Paula Magariños, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A. Patrícia Bento, Melissa F Adasme, Peter Monecke, Gregory A Landrum, and Andrew R Leach.
The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods.
Nucleic Acids Research, 52(D1), Nov 2023.