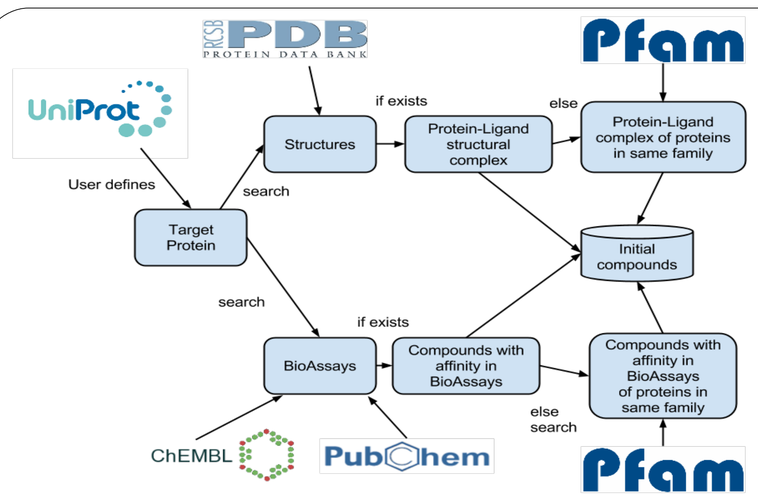
LigQ: An open-access protein & ligand structural similarity based tool for the enrichment of Virtual Screening compound sets

Leandro G. Radusky[1], Javier Luque[2], Xavier Barril[2], Marcelo A. Marti[1], Adrian G. Turjanski[1]

[1]Department of Biological Chemistry , FCEyN, University of Buenos Aires, Intendente Güiraldes 2160, C1428EGA, Buenos Aires, Argentina [2]Department of Physical Chemistry, Faculty of Pharmacy and Institute of Biomedicine (IBUB), University of Barcelona, Avgda. Diagonal 643, Barcelona 08028, Spain

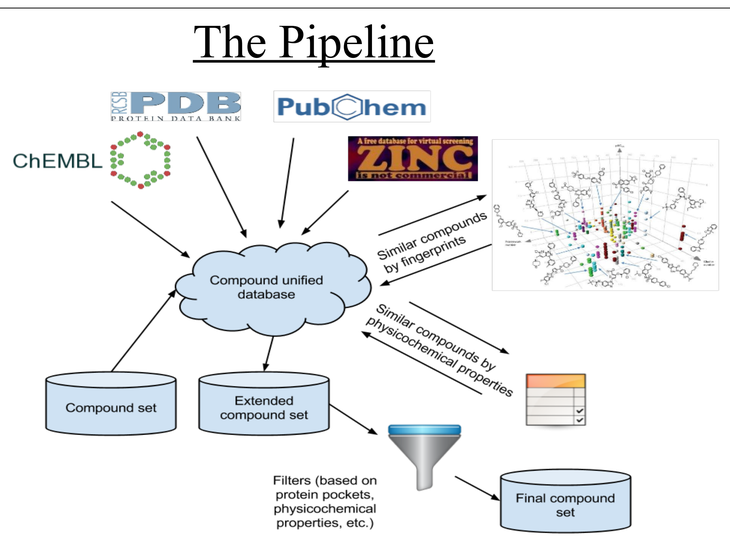
Abstract

The discovery of small molecules with biological activity, is of great interest for the development of therapeutic agents. One problem is how to narrow the search of possibly active compounds from large ligand databases. In this work we developed a bioinformatic pipeline which involves all steps from protein structure determination to ligand docking of the selected compounds against the protein target.

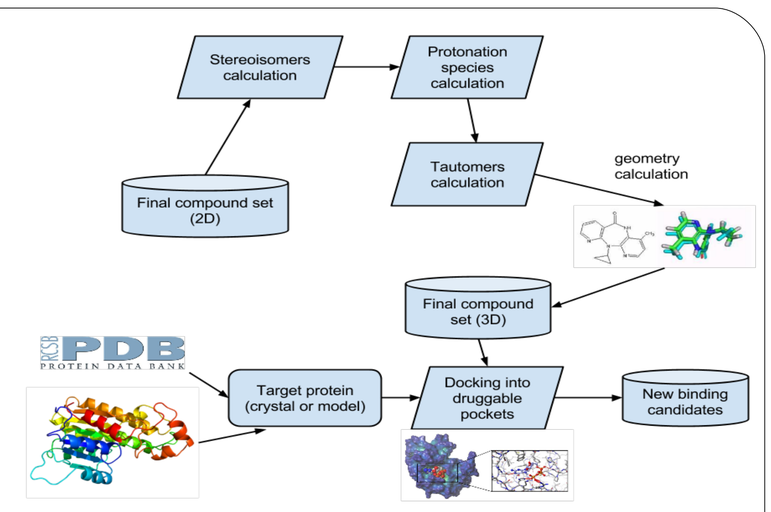


Initial compound determination: A protein (or a set of proteins belonging to a genome) are used as pipeline input(5). The set of initial compounds are fulfilled with the ligands within resolved structures of the protein(4), or optionally of the structures of the protein in the same protein family. Also, the drug-like compounds reported in bioAssays databases are used in the initial compound set.

The Pipeline



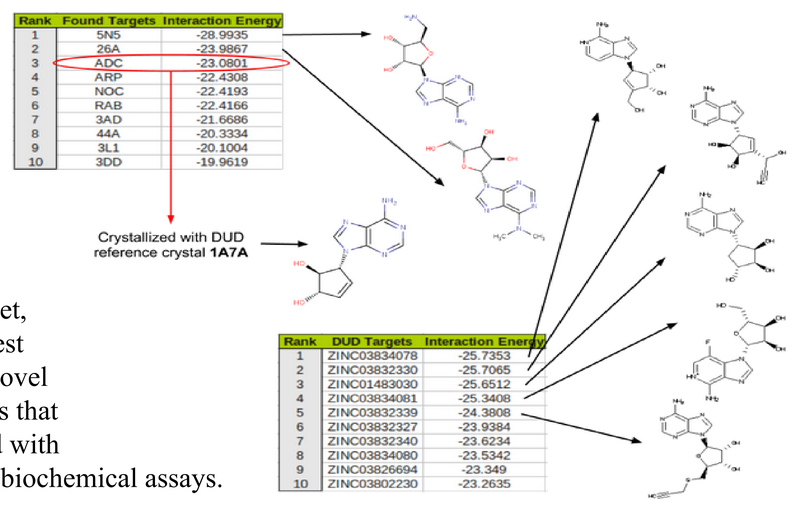
Compound Set Enrichment: The initial set is expanded with novel binding candidates, based in structural similarity (fingerprints) and physiochemical properties filters performed against large compound databases (1,2,3). Very similar compounds are clustered.



Geometry determination and Virtual Screening: For all the candidate compounds, calculations are performed to determine all his valid three dimensional structures(7). Docking (6) grid over protein structure plus packed candidate ligands are available for download and perform docking calculations with rDock or AutoDock software.

Discovering new binding candidates

Candidate compounds proposed by the method where docked against all the proteins belonging to the DUD targets dataset, and some of the best Ranked ones are novel Binding candidates that has to be validated with the correspondent biochemical assays.

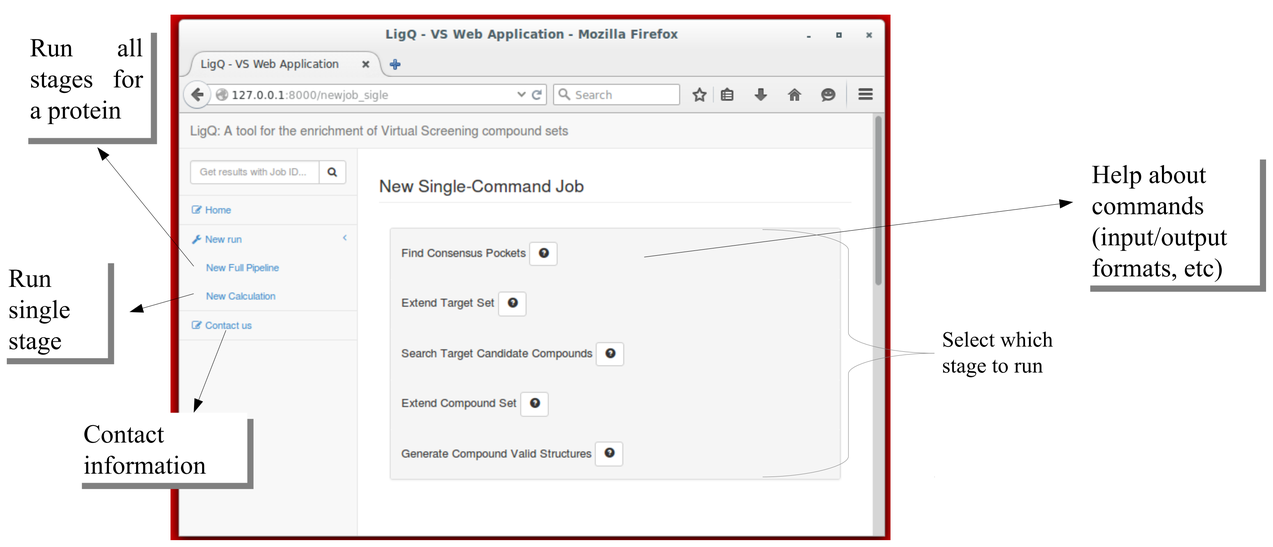


Rank	Found Targets	Interaction Energy
1	5N5	-28.9935
2	26A	-23.9867
3	ADC	-23.0801
4	ARP	-22.4308
5	NOC	-22.4193
6	RAB	-22.4166
7	3AD	-21.6686
8	44A	-20.3334
9	3L1	-20.1004
10	3DD	-19.9619

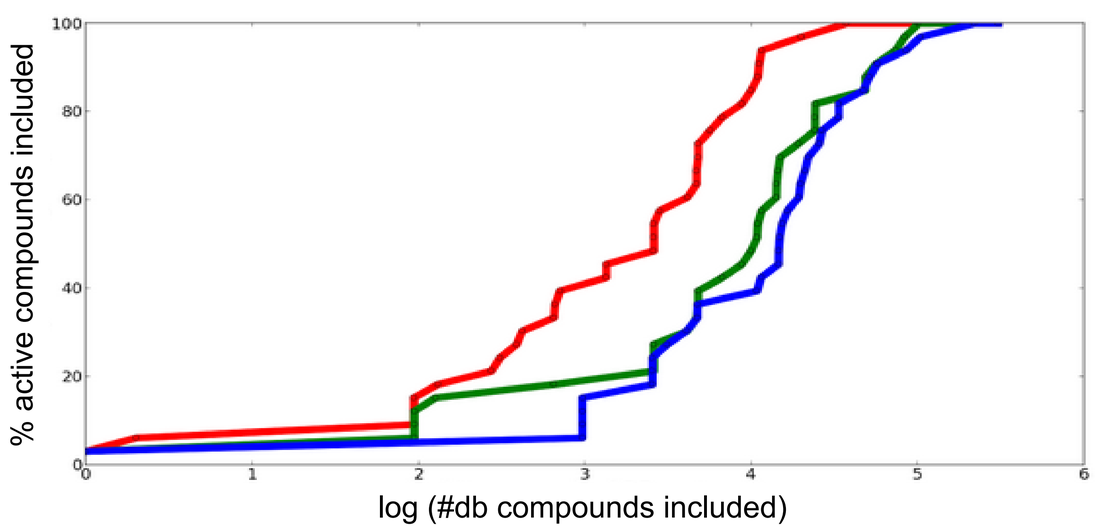
Crystallized with DUD reference crystal **1A7A**

Rank	DUD Targets	Interaction Energy
1	ZINC03834078	-25.7353
2	ZINC03832330	-25.7065
3	ZINC01483030	-25.6512
4	ZINC03834081	-25.3408
5	ZINC03832339	-24.3808
6	ZINC03832327	-23.9384
7	ZINC03832340	-23.6234
8	ZINC03834080	-23.5342
9	ZINC03826694	-23.349
10	ZINC03802230	-23.2635

Performing calculations in the Web Sever



Virtual screening set enrichment



Recovered Ligands of DUD	Crystallized target with drug (A)	Crystallized target without drug (B)	Only targets of the family (C)
33/33 (100%)	304205 (all)	304205 (all)	304205 (all)
33/33 (100%)	37431	101182	219298
30/33 (90%)	11165	56274	57340
25/33 (75%)	5658	24175	26626
17/33 (50%)	2592	10821	14806
9 (25%)	388	2588	3124

Having a curated database of non-redundant compounds of ~300k molecules, we move for a single protein with 33 active molecules the similarity Tanimoto factor in the pipeline, capturing one active molecule at a time.

For each step, we see how many molecules of the whole dataset were captured.

In the red case, we run the pipeline starting from a protein with a crystal having a drug coupled within. In the green case we keep a structure but remove from the db all the coupled to drug ones. In the blue case we only keep proteins of the same family not couplet.

We can observe that in the best case we can recover all the active original compounds having to screen only the ~10% of the original database, and the 90% of the compounds screening only the 4% of the database. With a less favorable seed the enrichment is worst but also good for enrich the screening dataset.

Conclusions

- We have developed a pipeline that allows the community to find better sets of compounds to perform virtual screening computations, giving to the users the option of download all the resources needed to perform the docking over the candidate binding region in-house.
- The presented tool is also useful to find in big compound databases novel molecules that have good chances to bind to the protein of interest of the user.
- We performed calculations over the proteins of the Database of Useful Decoys (DUD) having for all this targets similar set enrichment factors.
- For all this proteins we also obtained novel compounds with promising results based on docking energies that has to be confirmed as coupling ligands with posterior biochemical assays.

References

(1) Pubchem: Wang et. al. (2009) **(2)** ChEMBL: Gaulton et. al. (2012)
(3) ZINC: Irwin et. al. (2005) **(4)** PDB: Kouranov et. Al (2006)
(5) UniProt: UniProt Consorptium (2008) **(6)** rDock: Ruiz-Carmona et. Al (2014)
(7) Calculator Plugins: ChemAxon 15.6.29.0

Acknowledgements

Leandro Radusky is a PhD fellow of Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET). Adrián Turjanski y Marcelo Martí are CONICET researchers and University of Buenos Aires professors. Javier Luque and Xavier Barril are University of Barcelona Professors and ICREA researchers.