Dr James Gheen

Editorial Assistant JCIM

Reference ci-2017-00241w - LigQ: a WebServer to select and prepare ligands for virtual screening.

Please find below a revised version of the referenced manuscript with all changes are highlighted in red.

Best regards,

Leandro Radusky and Marcelo A. Martí

LigQ: a WebServer to select and prepare ligands for virtual screening

Leandro Radusky[1,2], Sergio Ruiz-Carmona[3], Carlos Modenuti[1,2], Xavier Barril[3,4], Adrian G Turjanski[1,2] and Marcelo A. Martí[1,2].

[1] Departamento de Química Biológica , Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

[2] Insituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN-CONICET), Pabellón II, Buenos Aires C1428EHA, Argentina

[3] Department of Physical Chemistry, Faculty of Pharmacy and Institute of Biomedicine (IBUB), University of Barcelona, Avgda. Diagonal 643, Barcelona 08028, Spain

[4] Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, Barcelona 08010, Spain

[*] Corresponding Author: marti.marcelo@gmail.com

Abstract

Virtual screening is a powerful methodology to search for new small molecule inhibitors against a desired molecular target. Usually, it involves evaluating thousand of compounds (derived from large databases) in order to select a set of potential binders that will be tested in the wet-lab. The number of tested compounds is directly proportional to the cost, and thus the best possible set of ligands is the one with the highest number of true binders, for the smallest possible compound set size. Therefore, methods that are able to trim down large universal datasets enriching them in potential binders are highly appreciated. Here we present LigQ, a free WebServer that is able to i) determine best structure and ligand binding pocket for a desired protein, ii) find known binders, as well as potential ligands known to bind to similar protein domains iii) most important, select a small set of commercial compounds enriched in potential binders and iv) prepare them for virtual screening. LigQ was tested with several proteins, showing an impressive capacity to retrieve true ligands from large datasets, achieving enrichment factors of over 10%. LigQ is available at http://ligq.qb.fcen.uba.ar/.

1. Introduction

The control of a protein's activity by means of small organic molecules (drug like) is not only the hallmark of molecular pharmacology, but also a powerful tool to study protein function in its biological context[1,2]. Therefore, strong efforts are devoted to the search of compounds that are able to fulfill this task. In-silico virtual screening (VS) is one of the most powerful and widely used methodologies to search for lead compounds, that bind to the protein of interest with moderate to high affinity. VS is commonly defined as the automated *in-silico* evaluation of very large libraries of compounds, in order to select potential binders that are further studied or directly tested in the wet-lab. VS can combine different methodologies, and they usually include a prefiltering step based on chemical properties, which can be followed by an automated docking protocol, and finally a more detailed estimation of the binding free energy[3].

Since the main cost of a VS project is directly related to the number of compounds to be tested experimentally and usually a relative low number (ca. hundred) of compounds are selected, it is very important that this ~~selected~~ set contains the highest number of true binders. Similarly, when prefiltering compounds from huge datasets to perform molecular docking, since the computational cost of the docking procedure is directly related to the number of compounds in the docking set, filtering methods that enrich docking sets in true binders, are highly desirable.

Moreover, compounds are usually derived from databases in some type of one dimensional format (SMILES, InChI, fingerprints, etc.) and several *in-silico* procedures are required to convert each of them to a proper 3D structure to perform the molecular docking, particularly if all relevant enantiomers and tautomers at physiological pH are to be considered[4]. Also, the

receptor protein structure has to be manipulated and the binding pocket, needs to be defined properly to improve the VS outcome[5]. Although, there are several free softwares able to perform these tasks, like for example Fpocket[6,13], DogSiteScorer[7], Openbabel[8] and SiteMap[26] among many others, most of them require local installation, cheminformatics skills, and are time consuming, which represents a drawback for their widespread use by experimentally focused groups.

LigQ, presented here, is to our knowledge the only free on-line web service that allows performance of fast VS procedures with an impressive capacity to detect potential binders to a desired target, which also allows their preparation for further screening using docking procedures. LigQ is organized in four independent modules that can be used sequentially to perform all VS preparation steps (see Scheme 1): i) the Pocket Detection Module (PDM), which allows users to find optimal the binding pocket for a given protein target, ii) the Ligand Detection Module (LDM), which allows users to find a seed group of potential binders known to modulate similar proteins, iii) the Extend Ligand Set (ELS) module, which allows (using a chemical similarity cut-off) to extend the list of potential binders from a large database of commercially available compounds, and finally, iv) the Ligand Structure Generation (LSG) module, which generates all relevant enantiormer/tautomer 3D structures for the desired ligands to use them in molecular docking. Together, these modules facilitate the process of *in-silico* drug discovery and increase the chances of finding active compounds against the desired target.

2. Computational Methods

Pocket Detection Module (PDM): The pocket detection module performs the task of finding best ligand binding pocket in the desired protein structure. The module takes as input a Protein Data Bank[10] (PDB) structure (pdbid) or a UniProt[11] accession. When the input is UniprotID, structural information is obtained from the corresponding online xml data. If there is more than one available crystal structure, the best resolution crystal with greater than 0.7 horizontal coverage is retrieved. If no structure is available the module tries to build an homology model using Modeller software[12], which is widely used in the community and has been shown an excellent performance in previous works from our group[21, 27]. To ensure the quality of the model, only those with GA341 score above 0.7, QMEAN between -2 and 2 and over 60% coverage of the protein sequence are retained. Once the structure is obtained pockets are determined and ranked according to their druggability score using the FPocket[6,13] software, chosen due to its performance in previous works from the group[6, 13, 21, 27] were we showed that it is able to correctly detect protein pockets and accurately assess their druggability.

Ligand Detection Module (LDM): This module performs a database search for potential binders to the desired protein (input). First, each protein domain, is assigned to a Pfam[14] (Protein Families database) domain. Then, the LDM looks for all structures of these domains in the PDB, and also for any match to these domains in the ChEMBL bioactivity database[15]. Ligands corresponding to the matching entries are retrieved and shown in the website both as figures or in 3D with JSMol[16] visualizer. Also, its structure in SDF or MOL2 format can be retrieved from the server. These compounds correspond to the seed group and are classified in four groups according to how they were retrieved (see results section). If no proteins of the Pfam family have reported bioactivities nor co-crystallized compounds the seed set will be empty.

Extend Ligand Set Module (ELS): The ELS module compares those compounds from the seed set specified by the user, against all compounds in the LigQ database (see below). Comparison is performed using chemical similarity criteria based on Tanimoto Index[17], which gives a similarity score between 0 (completely different) to 1 (completely equal), using the Open Babel package[8]. All ligands which surpass a user defined chemical similarity cut-off are retrieved. Also, additional filtering to the resulting set can be applied to the chemical properties (volume between two values, logP, charge, etc.).

Ligand Structure Generation Module (LSG): The LSG modules uses JChem Java library[9] first to build all possible enantiomers and second to compute the pKa of each ionizable center and select most probable tautomers at physiological pH. The resulting compound structures are made available to the user both in SDF and MOL2 formats.

The LigQ database: The LigQ database was built by selecting all those compounds belonging to the ZINC[18] database (version 12) having the "purchasable" tag in the record, clustering them by 0.95 similarity index, and taking a random representative compound, plus those compounds that are present in the PDB cocrystallized with a protein receptor, plus those compounds that are labeled as "active" in the ChEMBL database in a bioassay with any given protein target. Therefore, those ligands retrieved from the database are expected to be readily available for purchase and future experimental testing. All ligQ results pages, link each compound with all other similar purchasable ones (TI>0.9 in ZINC). The size of the LigQ database is ~1.5M compounds.

Further details on how to use each of the LigQ modules can be found in Supporting Information in the form of a user tutorial.

Performance metrics: To analyze the LigQ ELS performance we determined for each protein, seed set, and a few cut-off values, the enrichment factor (EF) defined as: EF(%)= (RTL/RC)*100. Where RTL is the number of retrieved true ligands and RC is the total number of retrieved compounds. The EF is thus defined as the probability of finding a true ligand among all the retrieved compounds, which in other words tries to measure the chance of finding a hit for a fixed set of compounds to be tested. The relative EF is simply the EF value for a given cut-off value, normalized by the EF computed for the whole database (i.e. the number of true ligands divided by the database size).

We defined four different seeds (sub)sets, based on the compounds source, which were considered in the tables and graphics of this work: Seed I, considering those ligands co-crystallized with the analyzed target; Seed II, considering those ligands co-crystallized with any target in the same Pfam family of the target; Seed III, considering ligands with a bioassay over the target; and Seed IV, considering ligands with a bioassay over any protein in the same PFam family of the target.

We also computed semilogarithmic-ROC curves (Figures 1 and S1) as a function of chemical-similarity cut-off. For this sake, true positives were defined as all known true ligands, excluding those found by the LDM and thus being part of the seed sets, that are retrieved by the ELS module. All other retrieved compounds are considered as false positives. True negatives are those compounds that are not true binders and are not retrieved, while false negatives are true

ligands which are lost (not retrieved) for a given cut-off value. In all cases the x-axis (False Positive Rate) is presented in logarithmic scale to make the plot easier to read.

Finally, dendrograms were built by computing the distance matrix between all pairs of compounds present in the seed set and the true ligand set, using SciPy[19] and Matplotlib[20].

3. Results

To show the potential of LigQ, we looked first at its capacity to detect true binders using the LDM, and secondly, at how the ELS is able to reduce the total number of retrieved compounds from the LigQ database, keeping most of the true binders as the chemical similarity cut-off is increased. The PDM and LSG have been already analyzed in other works and thus will be commented only briefly[21].

Analysis of LDM performance. It is difficult to test LigQ performance without involuntarily biasing the test for favorable cases, since as expected proteins that can be used as positive controls -those displaying a good number of known ligands- have usually been extensively studied and are thus highly annotated in the searched databases. Therefore, we first performed an analysis of the amount of available ligand information in the PDB and ChEMBL databases for the selected test cases. For nine proteins derived from the Directory of Useful Decoys (DUD)[22], we determined with the LDM the four seed sets described in the Computational Methods section.

The results presented in Table S1 show that different proteins (and their families) display a wide range of known ligands, from a few dozens to the many thousands (To analyze how the results are presented in the LigQ web interface please see the tutorial presented in SI). The number of

ligands derived from complex structures or bioassays for a unique protein are similar, except that for some cases, where there is no assay available. Extension to the whole Pfam family clearly increases the number of retrieved compounds between 2 to over 10 times, particularly when considering data derived from bioassays.

Analysis of ELS module performance. We now proceed to analyze the performance of the ELS module, for this sake we defined the seed set as either those ligands bound to the protein in the PDB (Seed I), those ligands reported to bind the protein in bioassays (Seed II), those ligands bound to any protein of the family in the PDB(Seed III) and those ligands reported to bind to any protein of the family in bioassays (Seed IV). In each case we used the ELS module to search the ligQ database, but adding all known true binders derived from the DUD database. It is important to note, that these true binders, which will be used a true positives in our analysis, are not retrieved by the LDM, since they are not reported, despite a few cases, in the PDB and ChEMBL databases. For each protein and seed set, we determined as a function of the chemical similarity cut-off, how many true binders and total compounds are retrieved. To analyze the results we built ROC curves as a function of chemical similarity cut-off and determined the corresponding EF as defined in methods. The results for two representative examples, are shown in figures 1 and S1, and summarized in Table 1.

The AMPc beta lactamase presents an ideal case (Figure 1, left). For the lowest tested chemical similarity cut-off (ca 0.5) required to retrieve 17 of the 21 true binders, using all seed compounds (red line), less than 10.000 total compounds are retrieved (EF=0.17%), which corresponds to a ca. 100 times increase in the EF (purple stars). Moreover, if the chemical similarity cut-off is increased to about 0.75, only hundred compounds are retrieved containing more than 10 true

binders, corresponding to an EF > 10%. Of course, and as expected, the high hit rate is counterweighted by the novelty and chemical diversity of the hits. As we increase the chemical similarity cut-off, although the EF decreases the active compounds differ more and more from the seed compounds. The plot also shows the effect of the seed set size. To retrieve, for example, about 14 true binders (black star), ca 10,000 total compounds are retrieved if only ligands derived from the PDB are considered as seeds (blue line); about 1,000 if ligands (red star) for all members of the protein family retrieved from bioassays are considered, and only 100 total compounds when all those ligands retrieved by the LDM are combined (cyan star). The results for AMPc beta lactamase thus show that with a medium cut-off the ELS module is able to significantly reduce the number of compounds that for example could be used in further docking experiments with high chances of detecting novel -different- compounds, while for higher cut-offs it already yields small enough enriched set of compounds to be directly used in wet lab experiments.

The results for DHFR, shown in Figure 1 (right), are not as impressive. In this case, no significant drop in the number of retrieved compounds from the LigQ database, while retaining most of true binders, is observed as the chemical similarity cut-off is increased. However, increasing the cut-off still results in an important increase in the EF. For example, for an initial 0.52 chemical similarity cut-off about 200 true binders in 14,000 compounds are retrieved (EF of 1.5%, purple star), while the EF increases to 10% (for a total of 1,200 retrieved compounds) for a chemical similarity cut-off of 0.72 (blue star).

To better understand the relationship between the diversity of the seed set, the number of recovered true binders, and the total number of retrieved compounds, we computed the chemical

similarity distance matrix for all ligands in the seed set together with the true binders, and built the corresponding dendrograms. The results, presented in Figure S2, show that for AmpC true binders are well distributed within the available seed ligands. Therefore, seed ligands are usually closer (in terms of chemical similarity) to all true binders, and when the cut-off in increased they remain in the retrieved set. On the contrary for DHFR, true binders cluster together, separately, and usually at moderate distances from the seed compounds. Thus, when high chemical similarity cut-offs are used, these clusters of true binders are dropped down. To include them lower cut-offs are required which in turn result in larger number of total retrieved compounds.

Beyond the selected examples, the summary of the results obtained for all test cases are presented in Table 1, where we selected only one representative protein for each Pfam family present in the DUD dataset. The table shows that while for about 10K total compounds, the percentage of true ligands is still small (EF =< 1%), values increase to more than 10% in some cases for a total of 100 retrieved compounds. These values represent a relative increase in the EF of 1 to 10 thousand times. Concerning the seed sets, although larger sets represent an increase in the EF, the increase is moderate. For the whole DUD dataset, LigQ achieves an average area under the ROC curve of 0.71 and an average enrichment factor at 1% of 17. These values are slightly lower (see table S3) compared to methods like LIGSIFT[23] of mRaise[24], and similar (or even higher) compared to other methods[25]. However, it is important to note, that unlike LigQ these are standalone tools that need to be installed locally, and combined with other tools if the user wants to reproduce our automatic pipeline.

Analysis of LGS module. Last but not least, we analyzed how the LSG Module increases the number of structures that need to be docked in relation to the number of selected compounds to

be tested experimentally. The results presented in Table S2 show that the number of structures is usually between 4 to 8 times the number of selected ligands, which therefore is an important point to consider.

4. Discussion

The aim of LigQ is to provide an easy to use and friendly web based application to aid researchers in the performance of Virtual Screening projects that look for potential binders to their protein of interest. We expect LigQ to be of potential interest first, to experimental groups working on proteins with small (or total lack of) experience in the use of *in-silico* tools to study protein-ligand interactions. And second, also to those groups that routinely perform VS projects, that could benefit from the straight forward, integrated, and potential increase in the success rate and knowledge provided by the LigQ modules, particularly the ELS.

The key, and novel, elements of LigQ are the LDM and ELS. The first, of these, allows user to rapidly determine not only known binders for a protein of interest, but those ligands known to bind to homologous protein domains, thus yielding a potential set of compounds to test experimentally (the seed set). Most important, application of the ELS allows performing a database search and using a chemical similarity cut-off criteria, retrieve a selected number of commercial available compounds which are significantly enriched in real binders. Depending of the amount of prior information available -i.e. the size and diversity of the seed set, the ELS can simply generate a high diversity set for future VS studies, or a achieve, as shown for AMPc, a double-digit hit rate at the top of the compound list. Of course, if there is a complete lack of information and no seed compounds are found by the LDM, other strategies can be employed to

nonetheless exploit the ELS potential. For example, seed set can be built by using more distantly related proteins, such as functionally related protein from another family, or ligands that bind to similar pockets in completely unrelated targets (using a pocket comparison tool).

Another future direction of potential LigQ improvement, concerns the chemical similarity criteria used by the ELS. LigQ current enrichment potential is manly related to its database search capacity and less to its chemical similarity comparison method. It is well known that different chemical comparison tools (or even flavors of Tanimoto like indexes) show significant different scores (and thus similarity criteria) for any pair of molecules. Particularly, as shown above, other methods used for ligand comparison and retrieval such as mRAISE or LIGSIFT show slightly better performance for finding real binders. However, LigQ has the advantage of performing also the pocket and known ligand detection steps, which are seamlessly integrated to ELS module where the ligand comparison is performed. Moreover, due its modular architecture, these other better ligand comparison methods could be easily joined to LigQ offering more flexible and possible better search possibilities, yielding a great potential to improve LigQ. Both of these ideas, which can be viewed as different modes of protein-protein and compound-compound comparison methods, will be analyzed and developed in future versions of LigQ.

5. Conclusions

We present a web based tool that significantly aids the researcher in the selection and preparation of ligands -as well as the desired protein receptor- for a Virtual Screening procedure. The application is divided in four semi-independent modules that allow the user to determine: i) the receptor ligand binding pocket, ii) the known ligands for the desired protein and family members

iii) a set of purchasable ligands enriched in potential binders and iv) 3D structures for selected ligands ready for the VS. Moreover, studies in a small set of well known proteins shows that the ELS module is able to yields EF from 0.1 to 10% for sets containing between 10,000 to only 100 compounds. This enrichment is key for either reducing the computational cost of subsequent docking experiments, or to directly test the retrieved compounds in the wet lab, maximizing the chance of success. In summary, LigQ allows users to balance the cost-benefit according to their needs. At one end of the spectrum, one may by-pass the docking step and test a few tenths of candidate compounds with good probability of finding hits. At the other end, one may dock all the compounds, without considering any previous knowledge. The typical situation we envisage is where the user decides to save 50-99% of calculation time, yet wants to maximize the hit rate.

Figures

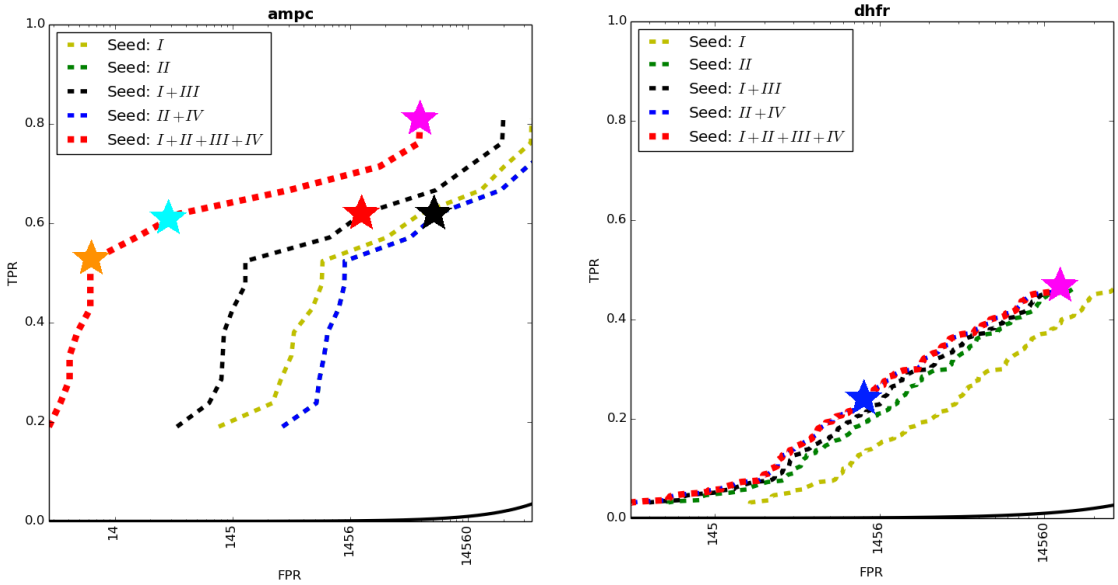
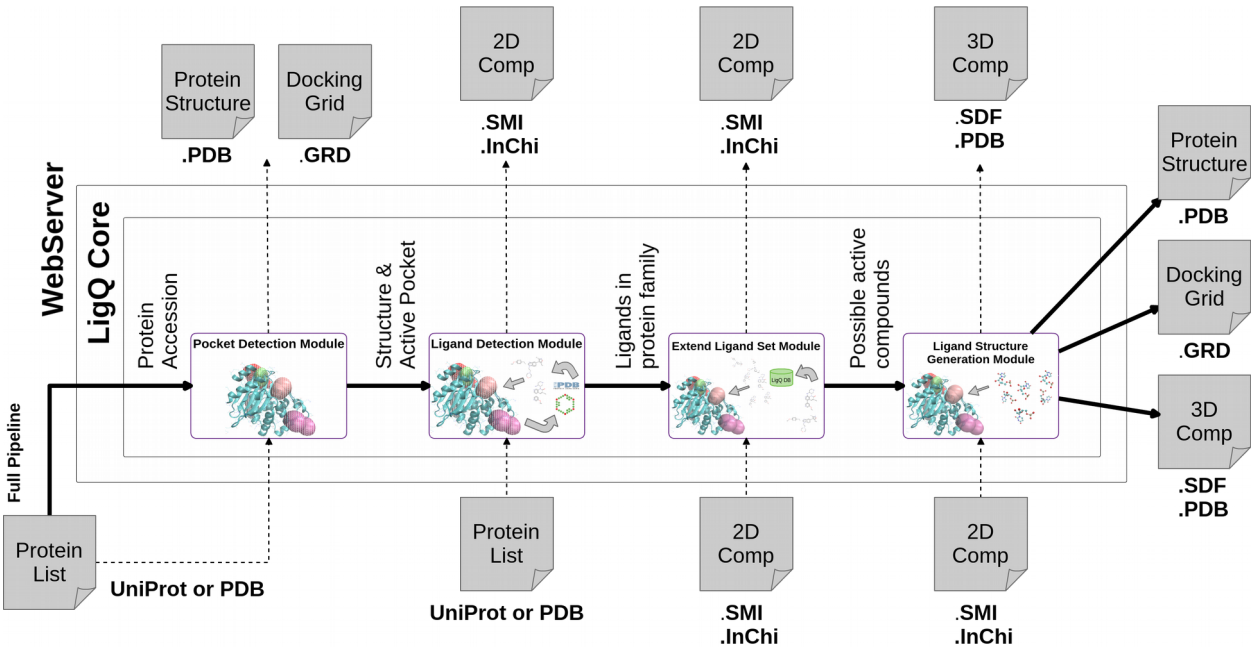


Figure 1. Plots of the semilogarithmic ROC curves for AmpcC (left) and DHFR (right). TPR is defined as the number of retrieved true binders relative to the total number of true binders. The FPR is defined as the number of total retrieved compounds with respect to the whole database size. FPR label indicates the actual number of retrieved compounds for clarity purposes. Green, blue, yellow and black correspond to seed sets I, II, III and IV, as described in Computational Mehods. For AmpC, green and blue lines superimpose. Red for all sets joined.

Schemes

Scheme 1. Flow Scheme of the LigQ WebServer Pipeline.



Tables

Protein Receptors[a]		**Seed Set I**			**Seed Set III**			**All seed Sets**		
	# RC_b	100	1K	10K	100	1K	10K	100	1K	10K
	# $TL(\downarrow)_c$	EF(%)(**Relative EF**)[d]			EF(%)(**Relative EF**)			EF(%)(**Relative EF**)		
ace	49	7 **2080**	1.3 **386**	0.2 **59**	8 **2377**	1.5 **445**	0.21 **62**	11 **3268**	1.5 **445**	0.21 **62**
ada	39	5 **1866**	0.8 **298**	0.16 **59**	7 **2613**	1.1 **410**	0.16 **59**	7 **2613**	1.3 **485**	0.16 **59**
ampc	21	4 **2773**	1.2 **832**	0.14 **97**	4 **2773**	0.9 **624**	0.14 **97**	14 **9706**	1.5 **1040**	0.17 **117**
dhfr	410	12 **426**	4.9 **174**	1.3 **48**	15 **532**	7.5 **266**	1.6 **58**	21 **745**	9 **319**	1.7 **61**
gart	40	2 **728**	0.5 **182**	0.21 **76**	7 **2548**	2.1 **764**	0.21 **76**	7 **2548**	2.1 **764**	0.21 **76**
gbp	52	12 **3360**	2 **560**	0.35 **98**	20 **5600**	3.2 **896**	0.43 **120**	20 **5600**	3.2 **896**	0.43 **120**
na	49	26 **7725**	4.3 **1277**	0.47 **139**	30 **8914**	4.4 **1307**	0.47 **139**	30 **8914**	4.6 **1366**	0.47 **139**
pnp	50	2 **582**	0.2 **58**	0.04 **11**	6 **1582**	1.8 495	0.2 **58**	8 **2329**	2 **582**	0.2 **58**
tk	22	12 **7941**	1.8 **1191**	0.22 **145**	13 **8603**	1.9 1257	0.22 **145**	13 **8603**	1.9 **1257**	0.22 **145**

Table 1. LigQ ELS performance module for all tested proteins. a) Abbreviations of the protein receptors names taken from the DUD, b) Total number of retrieved compounds for the selected cutt-off. For comparative purposes we fixed these values at 10 thousand, thousand and hundred

compounds. c) Total number of true ligands for the corresponding receptor. d) Enrichment factor (in %), and relative enrichment factor, as defined in Computational Methods.

Supporting Information

Supporting information available: Figures referred in the main text that supports the results and the validation of the tool, and a tutorial of how to use each LigQ module.

AUTHOR INFORMATION

Corresponding Author

* Marcelo A. Martí: marti.marcelo@gmail.com

Funding Sources

This work has been supported by grants PICT-GSK-2012-057 and PIP1220110100850 awarded to Marcelo A. Martí, by PICT-2010-2805 awarded to Adrián Turjanski and by the Spanish Ministerio de Economía (SAF2012-33481), the Catalan government (2014 SGR 1189) awarded to Xavier Barril.

Acknowledgments

The authors are thankful to Lucas A. Defelipe and Juan Pablo Arcón for testing the developed tool. We acknowledge ChemAxon[9] for the academic license agreement.

REFERENCES

1. Shoichet, B. K., Virtual screening of chemical libraries. *Nature* **2004**, 432, 862-865.

2. McInnes, C., Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.*, **2007,** 11, 494-502.

3. Ruiz-Carmona, S., Alvarez-Garcia, D., Foloppe, N., Garmendia-Doval, A. B., Juhos, S., Schmidtke, P., Barril, X., Hubbard, R.E., Morley, S. D., rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Comput. Biol.*, **2014,** 10, e1003571.

4. Barril X, Hubbard R.E., Morley S.D., Virtual screening in structure-based drug discovery. *Mini Rev. Med. Chem.*, **20014,** 4, 779–791.

5. Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., Sherman, W., Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.*, **2013,** 27, 221-234.

6. Le Guilloux, V., Schmidtke, P., Tuffery, P., Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, **2009,** 10, 168.

7. Volkamer, Andrea, Daniel Kuhn, Friedrich Rippmann, and Matthias Rarey. DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics,* **2012**, 28, 2074-2075.

8. OLBoyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., Hutchison, G. R., Open Babel: An open chemical toolbox. *J. Cheminf.*, **2011,** 3, 33.

9. Csizmadia, F., JChem: Java applets and modules supporting chemical database handling from web browsers. *J. Chem. Inf. Comput. Sci.*, **2000**, 40, 323-324.

10. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E., The protein data bank. *Nucleic Acids Res.*, **2000,** 28, 235-242.

11. UniProt Consortium, . The universal protein resource (UniProt). *Nucleic Acids Res.*, **2008,** 36, D190-D195.

12. Webb, B., Sali, A., Comparative protein structure modeling using Modeller. *Curr. Protoc Bioinformatics*, **2014**, 5-6.

13. Schmidtke, P., Barril, X., Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.*, **2010**, 53(15), 5858-5867.

14. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., , Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., Eddy, S. R.The Pfam protein families database. *Nucleic Acids Res.*, **2004**, 32, D138-D141.

15. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., , Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **2012**, 40, D1100-D1107.

16. Herraez, A., Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **2006**, 34, 255-261.

17. Bajusz, D., Rácz, A., Héberger, K., Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. *J. Cheminform.*, **2015**, 7, 1.

18. Irwin, J. J., Shoichet, B. K.. ZINC-a free database of commercially available compounds for virtual screening. J. Chem. Inf. Model., **2005**, 45, 177-182.

19. Jones, E., Oliphant, T., Peterson, P., SciPy: Open source scientific tools for Python, **2001**. URL: http://www. scipy.org, .

20. Hunter, J. D., Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **2007**, 9, 90-95.

21. Radusky, L., Defelipe, L. A., Lanzarotti, E., Luque, J., Barril, X., Marti, M. A., Turjanski, A. G., TuberQ: a Mycobacterium tuberculosis protein druggability database. *Database*, **2014**, bau035.

22. Huang, N., Shoichet, B. K., Irwin, J. J., Benchmarking sets for molecular docking. *J. Med. Chem.*, **2006**, 49, 6789-6801.

23. Roy, A., Skolnick, J., LIGSIFT: an open-source tool for ligand structural alignment and virtual screening. *Bioinformatics*, **2015**, 31, 539-544.

24. von Behren, M. M., Bietz, S., Nittinger, E., Rarey, M., mRAISE: an alternative algorithmic approach to ligand-based virtual screening. *J. Comput. Aided Mol*. Des., **2016**, 30, 583-594.

25. Kirchmair J, Distinto S, Markt P, Schuster D, Spitzer GM, Liedl KR, Wolber G, How to optimize shape-based virtual screening: choosing the right query and including chemical information. *J. Chem. Inf. Model.*, **2009**, 49, 678–692

26. Halgren, T. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, 49(2), 377−389

27. Radusky, L.G., Hassan, S.S., Lanzarotti, E., Tiwari, S., Jamal, S.B., Ali, J., Ali, A., Ferreira, R.S., Barh, D., Silva, A. and Turjanski, A.G. An integrated structural proteomics approach along the druggable genome of Corynebacterium pseudotuberculosis species for putative druggable targets. *BMC genomics*, **2015**, 16(5), S9.

FOR TABLE OF CONTENTS USE ONLY

LigQ: a server to select and prepare ligands for virtual screening

Leandro Radusky, Sergio Ruiz-Carmona, Carlos Modenutti, Xavier Barril, Adrian G Turjanski and Marcelo A. Martí.