Bioinformatics, 36(10), 2020, 3266–3267
doi: 10.1093/bioinformatics/btaa089
Advance Access Publication Date: 12 February 2020
Applications Note

Structural bioinformatics

InterLig: improved ligand-based virtual screening using topologically independent structural alignments

Claudio Mirabello and Björn Wallner *

Division of Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, Linköping 581 83, Sweden

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on August 14, 2019; revised on January 29, 2020; editorial decision on January 30, 2020; accepted on February 4, 2020

Abstract

Motivation: In the past few years, drug discovery processes have been relying more and more on computational methods to sift out the most promising molecules before time and resources are spent to test them in experimental settings. Whenever the protein target of a given disease is not known, it becomes fundamental to have accurate methods for ligand-based virtual screening, which compares known active molecules against vast libraries of candidate compounds. Recently, 3D-based similarity methods have been developed that are capable of scaffold hopping and to superimpose matching molecules.

Results: Here, we present InterLig, a new method for the comparison and superposition of small molecules using topologically independent alignments of atoms. We test InterLig on a standard benchmark and show that it compares favorably to the best currently available 3D methods.

Availability and implementation: The program is available from http://wallnerlab.org/InterLig.

Contact: bjorn.wallner@liu.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Virtual screening (VS) is a computational technique for the discovery of new, biologically active drug molecules. The idea behind VS is to analyze vast databases of libraries of untested compounds with *in silico* methods that should sift out the most promising leads before these are tested in experimental settings. Given the costs associated with laboratory experiments, it is no surprise that huge efforts are being put in developing more accurate methods for VS, so that fewer resources are wasted in pursuing potential dead ends. The two main approaches to VS are structure-based VS, where candidate ligands are docked on the structure of a known receptor, and ligand-based VS (LBVS), where the similarity of an active ligand is used to expand the number of potential candidate ligands.

LBVS methods are based on the assumption that structurally similar compounds have a higher chance of binding to the same receptor (Eckert and Bajorath, 2007). The structural similarity can be calculated by comparing 2D fingerprints of compounds or by more accurate 3D structural alignments (Hu *et al.*, 2018; Roy and Skolnick, 2015) that also allow scaffold hopping (Hu *et al.*, 2017) and provide starting points for 3D docking. Most methods using 3D are shape-based, representing molecules as a mixture of Gaussians and structure comparisons as overlaps between them (Roy and Skolnick, 2015), recently more detailed atom-level comparisons have shown promising results (Hu *et al.*, 2018).

In this study, we present InterLig, an open-source software for 3D-based LBVS. InterLig uses a simulated annealing-based procedure to map sets of atoms from two molecules in a topologically independent fashion, which makes it particularly suited for scaffold hopping. The simulated annealing procedure allows InterLig to compare tens of thousands of molecules within minutes (Supplementary Table S8). Moreover, along with the similarity score, a *P*-value is calculated to assess the statistical significance. InterLig is benchmarked against two state-of-the-art software for 3D LBVS and outperforms both according to several standard performance measures.

2 Datasets

InterLig is benchmarked against the Directory of Useful Decoys, Enhanced (DUD-E) (Mysinger *et al.*, 2012) containing 22 886 active ligands against 102 protein targets and 50 times more inactive decoys with similar physico-chemical properties but dissimilar 2D topology. For each protein target, a co-crystallized ligand is included in the set. In all tests, the 'seed ligand' is compared against active ligands and inactive decoys, and rank-based similarity to the seed ligand.

To account for the degrees of freedom of molecules, multiple conformers of DUD-E ligands and decoys are generated using OMEGA (Hawkins *et al.*, 2010) with the 'strict' flag set to false and minimum RMSD between two conformers set to 2 Å. Approximately 300k additional ligands and 8 M additional decoys are generated this way.

In addition, InterLig is also benchmarked against the Maximum Unbiased Validation (MUV; Rohrer and Baumann, 2009) set, which has been developed to correct for possible biases affecting the

© The Author(s) 2020. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/36/10/3266/5734641 by Linkopings universitetsbibliotek user on 13 August 2020

Table 1. Average enrichment factors (EFs) for different percentage of top hits and average area under the curve (AUC) for different sets

Set	Software	$EF^{1\%}$	$EF^{5\%}$	$EF^{10\%}$	AUC
DUD-E	LIGSIFT	16.88*	6.16*	3.95*	0.71*
	LS-align	20.70*	7.19*	4.44*	0.75*
	InterLig	**24.75**	**8.37**	**5.11**	**0.78**
DUD-E conformers	LIGSIFT	22.18*	7.48*	4.63 (0.07)	0.75 (0.12)
	LS-align	22.77 (0.07)	7.50*	4.63*	0.75 (0.07)
	InterLig	**23.79**	**8.03**	**4.88**	**0.77**
	InterLig + LS	**26.39**	**8.82**	**5.28**	**0.78**
MUV	LIGSIFT	4.31*	2.19*	1.75*	0.56*
	LS-align	3.15*	1.57*	1.27*	0.44*
	InterLig	**6.06**	**2.82**	**2.13**	**0.64**

Note: The highest values for each column are highlighted in bold.
*InterLig significantly ($P < 0.05$) better.

validation of LBVS methods. MUV includes 17 targets with 30 active ligands and 15 000 inactive decoys. Since no co-crystallized ligand is provided, each active is used as seed once and the result is the average of the metrics extracted for all 30 tests.

3 Results and discussion

InterLig is based on the InterComp algorithm that we recently developed and successfully applied to the comparison of protein interfaces (Mirabello and Wallner, 2018). It is capable of performing topologically independent alignments of sets of atoms in a 3D space while taking into account both the relative position and the chemical similarity of the aligned atoms. The core of algorithm in InterLig is identical to InterComp, the only difference is two parameters involving a cutoff distance, and the tradeoff between structural and chemical similarity. These parameters were decided on a set not used for testing the method (see Supplementary Material). The similarity measure used by InterLig depends on the size of the compounds, and smaller ligands have a higher probability to obtaining a high score by chance, the significance (*P*-value) of a score is calculated by fitting an extreme value distribution to scores for non-related ligands of different sizes (Supplementary Fig. S3).

InterLig is benchmarked using standard performance measures for VS (see Supplementary Material) against LS-align (Hu *et al.*, 2018) and LIGSIFT (Roy and Skolnick, 2015), two state-of-the-art software for 3D LBVS. LS-align has the best reported performance on the DUD-E benchmark, while LIGSIFT performs best on the older DUD set. To ensure a fair comparison each software was run using the default parameters on both the regular and the multiple conformer DUD-E benchmark. LS-align has a 'flexible' option to generate its own set of conformers, however when benchmarked it showed better performance in 'rigid' mode with the multiple conformers generated as above (Supplementary Table S7).

In the test on the regular DUD-E dataset, InterLig has significantly ($P < 0.05$) larger area under the curve (AUC) and enrichment factors (EFs) across all top rank percentages compared with both LIGSIFT and LS-align (Table 1). The performance metrics for LS-align are actually slightly better than those reported in the original publication (Hu *et al.*, 2018), most likely because how multiple compounds with the same ID are treated. Looking per target, InterLig has a higher AUC compared with LS-align for 60 and LIGSIFT for 76 (out of 102) targets (Supplementary Fig. S4a).

In the test on the regular MUV dataset, InterLig is significantly ($P < 0.05$) better than both LIGSIFT and LS-align on both EF and AUC (Table 1). Furthermore, we compare InterLig to other software that have been benchmarked on MUV set in another work (Tiikkainen *et al.*, 2009) and show that it outperforms them all (Supplementary Table S6). For the multiple conformers set, InterLig is significantly better compared with LIGSIFT and LS-align on most EF metrics but not on AUC, where it does perform better, but not significantly. Overall, InterLig seems to work slightly better when no conformers are considered. Detailed target-by-target results are available for regular (Supplementary Table S2) and multiple conformers (Supplementary Table S3). The results for multiple conformers are overall slightly better compared single conformers, indicating that it might be worth spending some additional time generating conformers to achieve optimal performance (Supplementary Fig. S5). However, the difference is not huge and if speed is of essence it is almost as good to not generate the conformers.

It was further noted that the per target performance for InterLig and LS-align were quite different (Supplementary Fig. S4). Thus, there should be potential to combine the two approaches to achieve even higher performance. To test this hypothesis, a combination of InterLig and LS-align was constructed by using the product of the reported p-values. Indeed, InterLig + LS-align is superior to both individual methods (Table 1), demonstrating that the results from the two methods are complementary.

Funding

This work was supported by a Swedish Research Council [2016-05369] the Swedish e-Science Research Center and the Foundation Blanceflor Boncompagni Ludovisi, née Bildt. The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the National Supercomputer Centre (NSC) in Linköping.

Conflict of Interest: none declared.

References

Eckert,H. and Bajorath,J. (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today*, **12**, 225–233.

Hawkins,P.C. *et al.* (2010) Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.*, **50**, 572–584.

Hu,J. *et al.* (2018) LS-align: an atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening. *Bioinformatics*, **1**, 10.

Hu,Y. *et al.* (2017) Recent advances in scaffold hopping: miniperspective. *J. Med. Chem.*, **60**, 1238–1246.

Mirabello,C. and Wallner,B. (2018) Topology independent structural matching discovers novel templates for protein interfaces. *Bioinformatics*, **34**, i787–i794.

Mysinger,M.M. *et al.* (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.*, **55**, 6582–6594.

Rohrer,S.G. and Baumann,K. (2009) Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.*, **49**, 169–184.

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

Tiikkainen,P. *et al.* (2009) Critical comparison of virtual screening methods against the MUV data set. *J. Chem. Inf. Model.*, **49**, 2168–2178.

Downloaded from https://academic.oup.com/bioinformatics/article/36/10/3266/5734641 by Linkopings universitetsbibliotek user on 13 August 2020