Application Notes

LightDock goes information-driven

Jorge Roel-Touris[1], Alexandre M.J.J. Bonvin[1] and Brian Jiménez-García[1*]

[1]Bijvoet Center for Biomolecular Research, Faculty of Science-Chemistry, Utrecht University, Utrecht, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The use of experimental information has been demonstrated to increase the success rate of computational macromolecular docking. Many methods use information to post-filter the simulation output while others drive the simulation based on experimental restraints, which can become problematic for more complex scenarios such as multiple binding interfaces.

Results: We present a novel method for including interface information into protein docking simulations within the LightDock framework. Prior to the simulation, irrelevant regions from the receptor are excluded for sampling (filter of initial swarms) and initial ligand poses are pre-oriented based on ligand input information. We demonstrate the applicability of this approach on the new 55 cases of the Protein-Protein Docking Benchmark 5, using different amounts of information. Even with incomplete or incorrect information, a significant improvement in performance is obtained compared to blind *ab initio* docking.

Availability: The software is supported and freely available from https://github.com/brianjimenez/lightdock and analysis data from https://github.com/brianjimenez/lightdock_bm5.

Contact: b.jimenezgarcia@uu.nl

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

1 Introduction

Computational tools are essential to predict and describe three-dimensional (3D) interactions between biomolecules. In particular, integrative approaches, i.e. data- or information-driven, are broadly used in order to combine experimental data with docking simulations (De Vries *et al.*, 2010, 2015; Russel *et al.*, 2012; Jiménez-García *et al.*, 2013; Rodrigues and Bonvin, 2014; Quignot *et al.*, 2018). In the context of molecular docking, there are still two main challenges: (1) searching the conformational space, especially in the case of highly flexible molecules, and (2) evaluating and selecting near-native poses out of the generated conformers, which is usually referred to as scoring.

LightDock (Jiménez-García *et al.*, 2018) is a multiscale flexible framework for the 3D determination of binary protein complexes based on the Glowworm Swarm Optimization (GSO) (Krishnanand and Ghose, 2009) algorithm that systematically optimizes the generated docking poses towards those energetically more favourable at every simulation step.

Introducing restraints or biases in docking is a powerful mechanism to drive the simulation towards poses that satisfy those restraints (Dominguez *et al*, 2003). Here we describe and benchmark an updated implementation of LightDock that now supports the use of information to drive or bias the docking simulation by filtering out swarms, pre-orienting ligand poses based on the available information and biasing the scoring energy upon satisfied residue contact restraints.

The results on the benchmark demonstrate a high performance of LightDock when used in combination with additional information. We also explore different scenarios with less accurate or incorrect information to show the versatility and robustness of our approach.

2 Methods

Due to the nature of the LightDock framework, information about interfacial residues can be applied at different levels depending on the availability of information for the receptor, the ligand or both. On the receptor side, we filter out initial swarms that are not in the proximity of the defined restraints (Suppl. S1), with the collateral advantage of reducing considerably the computation time. On the ligand side, we orient initial poses based on randomly selected receptor-ligand restraint pairs (Suppl. S2). Steps S1 and S2 are only performed at the initial setup stage of the simulation.

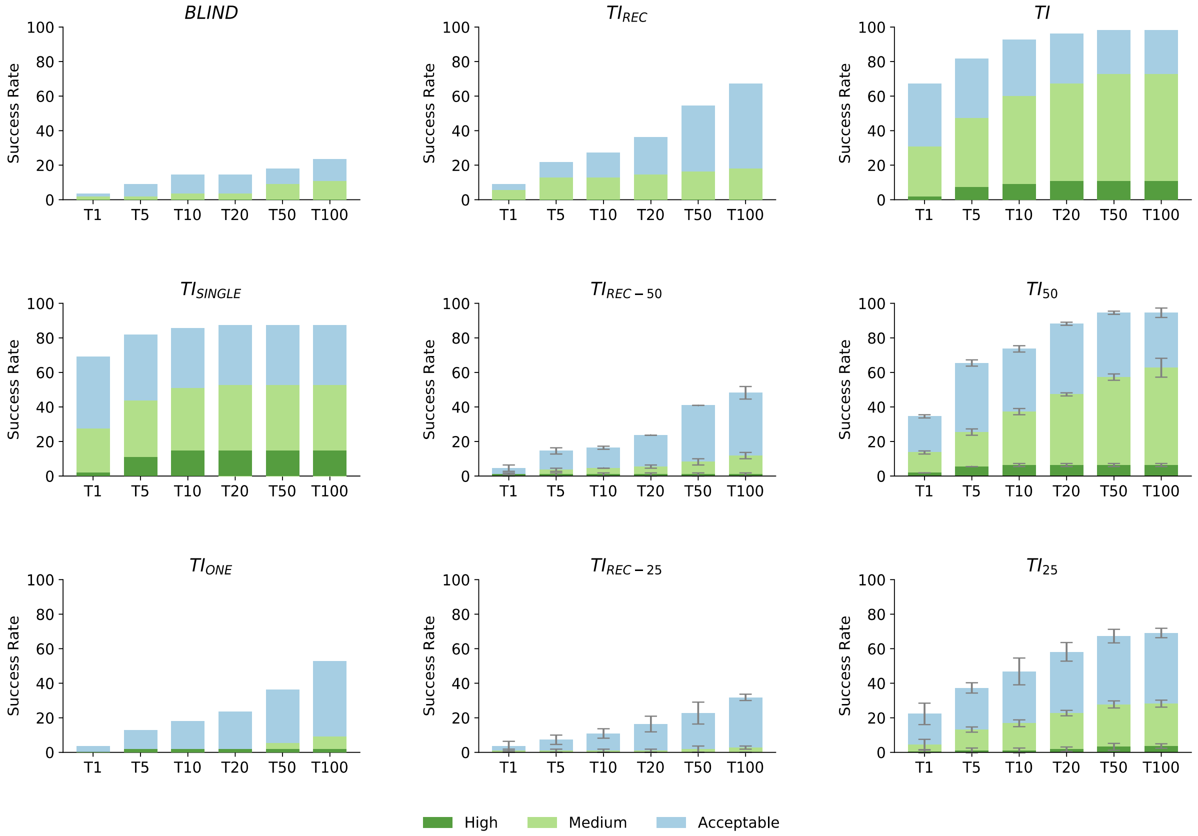

Fig. 1. Performance of LightDock for the nine different scenarios. *BLIND*: *Ab initio* docking. TI_{REC}: Only receptor contribution to the true interface. *TI*: All the residues from the true interface. TI_{SINGLE}: A single residue pair from the true interface. TI_{REC-50}: Half of the TI_{REC} and equal number of non-interfacial residues. TI_{50}: Half of the *TI* and equal number of non-interfacial residues. TI_{ONE}: Only one residue on the receptor, as defined in TI_{SINGLE}, is considered as restraint (i.e. no information on the ligand side). TI_{REC-25}: One fourth of the TI_{REC} and three times more non-interfacial residues. TI_{25}: One fourth of the *TI* and three times more non-interfacial residues. True interface residues are calculated at a cutoff distance of 3.9Å. Results are presented according to the CAPRI quality criteria (Lensink and Wodak, 2010) and the success rate is defined as the percentage of cases with at least one non-incorrect model within a given Top *N* (*N*=1, 5, 10, 20, 50 100).

Finally, we bias the scoring according to the percentage of satisfied residue contact restraints (Suppl. S3) at every simulation step. The biasing of specific residues could be disabled if they are defined as passive by the user (only S1 and S2 steps at the setup stage will therefore apply).

3 Results

The latest release of LightDock (0.7.0) (Jiménez-García *et al.*, 2019), which now supports the use of information to drive the docking in the format of residue restraints, was tested on the 55 unbound new entries of the Protein Docking Benchmark version 5 (Vreven *et al.*, 2015), which represents an unbiased dataset where no software/scoring functions were trained in, and includes 16 antibody-antigen complexes. We defined various scenarios to demonstrate its versatility and robustness as follows:

(1) *TI*: True interface, defined as those residues at 3.9Å distance (as also defined in LIGPLOT (Wallace *et al.*, 1995) by default) from the partner molecule. This is an ideal case where a fully accurate definition of interface residues is available, but no specific contacts are defined.

(2) TI_{50}: We defined two different artificial interfaces with half of the *TI* residues and equal number of non-interfacial residues forming a contiguous patch as described in Suppl. S4. Results are reported as averaged success rates of both runs (using each of the designed interfaces).

(3) TI_{25}: In the same way as in TI_{50}, we defined four different sets of restraints with one fourth of the original *TI* and three times more false positive residues forming a contiguous patch (Suppl. S4). Results are reported as averaged success rates of the four docking calculations (each one using a different artificially designed interface).

(4) TI_{REC}: Only the *TI* from the receptor is considered as restraints.

(5) TI_{REC-50}: As in TI_{50}, but only considering the receptor interface residues.

(6) TI_{REC-25}: As in TI_{25}, but only considering the receptor interface residues.

(7) TI_{SINGLE}: Only one receptor-ligand residue pair, making a real contact, is used as residue restraints.

(8) TI_{ONE}: Only one residue on the receptor, the same one as defined in $\mathrm{TI}_{\mathrm{SINGLE}}$, is considered as restraint, without any information on the ligand side.

While several docking algorithms allow the use of information as *a posteriori* filter, LightDock incorporates this data *a priori*. If residue restraints are provided, irrelevant sampling regions are excluded by filtering the initial swarms and pre-orienting the initial poses (glowworms). This method not only represents a more efficient way as compared to post-docking approaches but also leads to a higher success rate. To test this hypothesis, we have filtered the *BLIND* predictions (*$BLIND_{filtered}$*) according to an accurate description of the interface (residue restraints as used in *TI*). As shown in Suppl. Fig. 2, post-filtering results in a clear improvement of the performance compared to *ab initio* docking. Nevertheless, when using this information prior the docking (*TI*), the success rate considerably increases reaching a maximum of 98.2% for the Top50 (54 of 55 cases) compared to a moderate 40% in $\text{BLIND}_{\text{filtered}}$.

Figure 1 shows the results for the eight scenarios described above together with *ab initio* docking, which is included as a baseline for comparison purposes. The scoring function used in these LightDock simulations is DFIRE (Zhou and Zhou, 2009). When no prior information about the binding site is used for the docking calculations (*BLIND*), the predictive performance of LightDock lags behind any of the other scenarios tested in this work, with a moderate 14.5% and 23.6% success rates for Top10 and Top100 respectively. Interestingly, with the gradual use of information in the form of residue contact restraints, we find a boost in the performance up to a 92.7% for the Top10 when an accurate description of the interface (*TI*) is used. This represents an ideal case and illustrates how docking approaches can enormously benefit from integrating experimental data in their calculations. Unfortunately, structural experimental techniques rarely describe interfaces in a very accurate manner and the data produced is usually incomplete and/or incorrect, fact that heavily affect the performance of modelling approaches as previously discussed in Rodrigues and Bonvin, 2014.

To account for inaccurate or incorrect data, we have designed artificial interfaces with false positive residues (Suppl. S4). When only 50% of the original *TI* is used (*TI_{50}*) or 25% (*TI_{25}*), which represents 50% and 75% of non-interfacial residues, LightDock performance in Top10 is of 72.7% and 46.4% respectively. In the case of *TI_{50}*, Top100 performance compares to *TI* (94.6% vs 98.2%). This indicates that even when the information used to restrain the docking simulations in LightDock is incomplete and partially wrong, the protocol seems robust enough and still yields correct solutions for most of the cases (52 out of 55). However, the scoring becomes problematic compared to *TI* as the Top1 success rate drops from 65.5% to 33.6%.

In the scenario where only the contribution of the receptor is taken into account (*TI_{REC}*), a substantial success rate of 67.3% is obtained for the Top100. This scenario is especially interesting since it directly applies, for example, to antibody-antigen docking where no information about the epitope is known so the docking is performed exploring the whole surface of the antigen while for the antibody the HV loops are provided (Suppl. Figure 3). Moreover, when false positives are included in the *TI_{REC}* scenario (50% in *TI_{REC-50}*, 75% in *TI_{REC-25}*) the performance drops, but Top100 is still higher (46.3% and 28.2%) than *BLIND* (23.6%).

Finally, we push the limits of the algorithm defining only one residue restraint on the receptor molecule (this would mimic one mutation data point for example). This effectively means that, as in *TI_{SINGLE}*, only the ten closest swarms to the restraint will be generated, each of them containing randomly oriented glowworm poses (200 by default). In this scenario, restricting the sampling area helps the identification of near-native models as the performance is significantly higher than *BLIND* (Fig. 1). Remarkably, when we include a residue on the ligand molecule (*TI_{SINGLE}*), which is used in the pre-orienting step (Suppl. S2), LightDock predicts and scores a near-native solution in the Top1 for 69% of the cases. From the different tested scenarios, it seems reasonable to state that our protocol enormously benefits from the additional data in form of residue restraints, even when it is incomplete and/or partially incorrect.

4 Conclusion

The new version of LightDock offers a powerful tool for modelling protein-protein complexes with high accuracy when good quality information about interfaces is available. Next to enabling the incorporation of data from mutagenesis and/or bioinformatics predictions, for example, this strategy might also be convenient in scenarios such as limiting the sampling to the solvent accessible loops of a transmembrane protein, or the CDR loops of an antibody. Moreover, when incorrect and/or incomplete data are used to restraint the simulation, LightDock is still robust enough to yield valuable predictions. While other FFT-based methods do support *a posteriori* filtering, the pre-filtering of swarms in LightDock does lead to a reduction of the computation time and a higher performance, which could be used to ensure a denser sampling around the binding region.

Funding

This work has been done with the financial support of the European Union Horizon 2020 projects BioExcel (675728, 823830) and EOSC-hub (777536) and of the Dutch Foundation for Scientific Research (NWO) (TOP-PUNT grant 718.015.001).

Conflict of Interest: none declared.

References

Dominguez,C. *et al.* (2003) HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **125**, 1731–1737.

Jiménez-García,B. *et al.* (2019) brianjimenez/lightdock: Release 0.7.0. doi.org/10.5281/zenodo.3228412

Jiménez-García,B. *et al.* (2018) LightDock: A new multi-scale approach to protein-protein docking. *Bioinformatics*, **34**, 49-55.

Jiménez-García,B. *et al.* (2013) pyDockWEB: A web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. *Bioinformatics*, **29**, 1698-1699.

Krishnanand,K.N. and Ghose,D. (2009) Glowworm swarm optimization for simultaneous capture of multiple local optima of multimodal functions. *Swarm Intell.*, **3**, 87–124.

Lensink,M.F. and Wodak,S.J. (2010) Docking and scoring protein interactions: CAPRI 2009. *Proteins Struct. Funct. Bioinforma.*, **78**, 3073–3084.

Quignot,C. *et al.* (2018) InterEvDock2: An expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. *Nucleic Acids Res.,* **46***,* 408-416.

Rodrigues,J.P.G.L.M. and Bonvin,A.M.J.J. (2014) Integrative Computational Modelling of Protein Interactions. *FEBS J.*, **281**, 1998-2003.

Russel,D. *et al.* (2012) Putting the pieces together: Integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.*, **10**, e1001244.

Vreven,T. *et al.* (2015) Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.*, **427**, 3031–3041.

De Vries,S.J. *et al.* (2015) A web interface for easy flexible protein-protein docking with ATTRACT. *Biophys. J.*, **108**, 462-465.

De Vries,S.J. *et al.* (2010) The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.*, **5**, 883–897.

Wallace,A.C. *et al.* (1995) Ligplot: A program to generate schematic diagrams of protein-ligand interactions. *Protein Eng. Des. Sel.*, **8**, 127-134.

Zhou,H. and Zhou,Y. (2009) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714-2726.