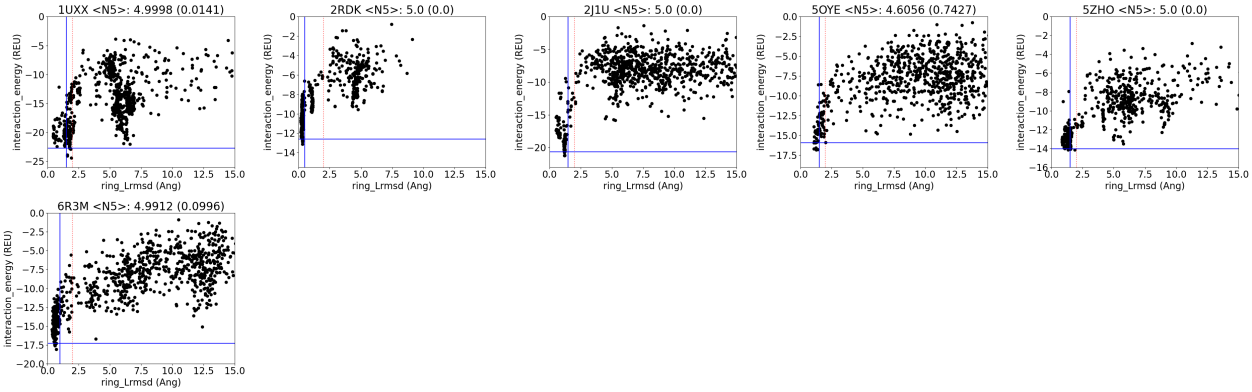
Scientific test: glycan_dock

FAILURES

None

RESULTS



AUTHOR AND DATE

Morgan L. Nance (morganlnance@gmail.com; June 2021)

Relevant publication

DOI: 10.1021/acs.jpcb.1c00910

Development and Evaluation of GlycanDock: A Proteinâ€“Glycoligand Docking Refinement Algorithm in Rosetta

Morgan L. Nance, Jason W. Labonte, Jared Adolf-Bryfogle, and Jeffrey J. Gray

The Journal of Physical Chemistry B 2021 125 (25), 6807-6820

-When "benchmark paper" is mentioned, it is referring to this publication

PURPOSE OF THE TEST

This test ensures that the GlycanDock protein-glycoligand docking refinement algorithm performs, at a minimum, on par with its observed performance during its original benchmarking. This test should catch if GlycanDock performance falls on a selection of 6 targets from the benchmark set.

BENCHMARK DATASET

This test uses 6 of the original 65 unbound protein-glycoligand targets from the benchmark paper. The unbound protein targets (opposed to the bound protein-glycoligand crystals) are used in this

test to provide a more accurate measure of GlycanDock performance in realistic docking refinement scenarios.

Input protein-glycoligand structures were all originally used as input for the GlycanDock benchmark paper. The constraints were also used in the benchmark (though here, for simplicity, each constraint used for each carbohydrate unit of the glycoligand are combined into a single .cst file). The reference native structures were also those used in the benchmark paper.

The unbound protein structure was aligned onto the bound protein structure using PyMOL. The coordinates for the glycoligand from the bound structure and the coordinates of the unbound protein structure were kept (to serve as the "native" unbound protein-glycoligand structure). This structure was then pre-packed. Finally, the glycoligand was randomly but systematically perturbed in rigid-body and glycosidic torsion angle space to reach the desired 7 +/- 0.1 Ang initial ring-RMSD.

The following 6 targets were selected for scientific benchmarking. See key results for detailed information on how the 6 targets performed in the benchmark paper.

*1UXX (unbound protein 1GMM)

-- CtCBM6 - xylopentaose

-- Reason for selection: Passed bootstrap_N5 >= 3 cutoff in benchmark but with relatively high sstandard deviation. CBM target represents a common protein of interest in glycobiology.

*2RDK (unbound protein 2Z21)

-- Cyanovirin - N-dimannose

-- Reason for selection: Performed perfectly in benchmark. Cyanovirin is an anti-viral glyco-binder of scientific interest.

*2J1U (unbound protein 2J1R)

-- Strep lectin - Blood Group A-tetrasaccharide

-- Reason for selection: Strong success with a glycoligand that is an important blood group antigen and is branched

*5OYE (unbound protein 5OYC)

-- CjGH5 - xyloglucan

-- Reason for selection: A difficult target (a hexasaccharide with two exocyclic branch points) that showed promise (bootstrap_N5 close to 3). This target serves as a potential indicator if Rosetta sampling and/or scoring improves

*5ZHO (unbound protein 5ZHG)

-- G4P RVC VP8* - A histo-blood group

-- Reason for selection: Strong success with a glycoligand that is an important blood group antigen and is branched. The human group C rotavirus VP8* protein is also an important target

*6R3M (unbound protein 1V0A)

-- CtCBM11 - beta-1,3-1,4-mixed-linked tetrasaccharide

-- Reason for selection: Just above the bootstrap_N5 cutoff for docking success, but with significant standard deviation. Glycoligand is of mixed linkage and serves as an interesting target for docking and glycosidic torsion angle sampling

PROTOCOL

The default GlycanDock docking refinement algorithm is applied to each input structure. Briefly, a small, random rigid-body rotation and perturbation is applied to the glycoligand's center-of-mass, and a small, random uniform perturbation is applied to each glycosidic torsion angle. The Metropolis criterion is not applied. Then, inner cycles of rigid-body and glycosidic torsion angle sampling occur while outer cycles control ramping the fa_atr and fa_rep score terms. The structure is minimized after every other sampling move, whereas packing occurs after every sampling move.

See the Methods section and Supplemental of the benchmark paper for more information on the GlycanDock algorithm

For the code, see src/protocols/glycan_docking/*

Approximately 1100 CPU hours

1UXX: 495 +/- 35 sec per model

2RDK: 358 +/- 19 sec per model

2J1U: 608 +/- 49 sec per model

5OYE: 1204 +/- 199 sec per model

5ZHO: 579 +/- 41 sec per model

6R3M: 642 +/- 153 sec per model

1000 models per target

((avg seconds * nstruct) * ...) / (60 sec * 60 minutes to convert to hours)

--> ((495*1000)+(358*1000)+(608*1000)+(1204*1000)+(579*1000)+(642*1000))/3600.

PERFORMANCE METRICS

We evaluate performance using the average N5 after bootstrap case resampling (bootstrap_N5). N5 means the number of near-native models (sub 2 Ang ring-RMSD) ranked within the 5-top-scoring. Ideal N5 is 5, making ideal bootstrap_N5 = 5.0. An N5 or bootstrap_N5 of 0 means that NO near-native models were ranked within the top-5.

Bootstrap case resampling is performed as follows. For a given target, 5000 sets of resampled models are generated by randomly selecting 1000 models with replacement from the original set

of models. The subset of models are then ranked by score (here, interaction_energy), and the N5 is determined. This process gives us 5000 N5 values per target, the average of which is our bootstrap_N5 value.

bootstrap_N5 >= 3 is considered a docking success. This was the performance cutoff used in the benchmark paper. For this scientific test, the cutoff for bootstrap_N5 differs per target.

This test also tracks if at least one of the 5-top-scoring models has a ring-RMSD (ring_Lrmsd) that is below the specified cutoff per target. These ring-RMSD values also come from the benchmark paper. However, the general cutoff for a near-native model is < 2.0 Ang ring-RMSD. The ring-RMSD metric check is not a pass/fail metric. ring-RMSD (ring_Lrmsd) is defined as the root-mean-squared deviation of all carbohydrate ring atoms of the glycoligand compared to the native/reference structure.

The following must be true for this scientific test to pass:

*1UXX: bootstrap_N5 >= 4

*2RDK: bootstrap_N5 >= 5

*2J1U: bootstrap_N5 >= 4

*5OYE: bootstrap_N5 >= 2 (Note: benchmark paper defined success as bootstrap_N5 >= 3)

*5ZHO: bootstrap_N5 >= 4

*6R3M: bootstrap_N5 >= 3 (Note: possibly too stringent given its bootstrap_N5 std dev (see above))

The cutoffs chosen are based on the target's performance in the benchmark.

The x-axis is the ring-RMSD (ring_Lrmsd) in Angstroms. The y-axis is the interaction energy (REU). Interaction energy is calculated by taking the total_score of the complex and subtracting off the total_score of the unbound complex (i.e. bound energy - unbound energy). One simulates the unbound complex by translating the glycoligand 1000 Ang away from the protein receptor.

The blue horizontal line marks the interaction_energy of the top-5 scoring model (i.e. the interaction_energy cutoff for the 5-top-scoring models). It is NOT a pass/fail metric.

The blue vertial line marks the ring_Lrmsd cutoff for that particular target (ranges from 0.5 to 1.5). It is NOT a pass/fail metric.

The red dashed vertical line marks the 2 Ang ring_Lrmsd cutoff for a model to be considered near-native (a general and universal model quality cutoff for protein-glycoligand complexes). It is involved in the calculation of bootstrap_N5, which is a pass/fail metric.

The title of each plot states the target ID, the bootstrap_N5 value and the standard deviation in parentheses.

KEY RESULTS

Results are compared to past performance as measured in the initial benchmarking of GlycanDock. Measuring performance requires the experimental structure of the protein-glycoligand complex, which is used here.

The following docking results are from the GlycanDock benchmark paper (NOTE, where nstruct was 2k; see the unbound benchmark set at initial ring-RMSD 7 +/- 0.1 Ang). Additional result details to look for in this scientific test are also given (beyond ensuring that all specified bootstrap_N5 cutoffs per target are hit/surpassed):

*1UXX (unbound protein 1GMM)

-- CtCBM6 - xylopentaose

-- # near-natives: 73

-- N5: 5

-- bootstrap_N5: 4.023

-- bootstrap_N5 std: 0.929

-- Future improvement desired: Better bootstrap_N5 discrimination (i.e. closer to 5)

-- What to look for: a good funnel toward near-native models with a secondary but less favorable funnel toward non-native-like models around 6 Ang

*2RDK (unbound protein 2Z21)

-- Cyanovirin - N-dimannose

-- # near-natives: 70

-- N5: 5

-- bootstrap_N5: 5.000

-- bootstrap_N5 std: 0.000

-- Future improvement desired: More near-native models generated

-- What to look for: A very strong funnel toward sub-Angstrom models

*2J1U (unbound protein 2J1R)

-- Strep lectin - Blood Group A-tetrasaccharide

-- # near-natives: 31

-- N5: 5

-- bootstrap_N5: 4.658

-- bootstrap_N5 std: 0.693

-- Future improvement desired: More near-native models

-- What to look for: A strong funnel toward near-native models, but with a comparatively large cloud of non-native-like models scored worse than the native-like ones

*5OYE (unbound protein 5OYC)

-- CjGH5 - xyloglucan

-- # near-natives: 13

-- N5: 5

-- bootstrap_N5: 2.665

-- bootstrap_N5 std: 1.357

-- Future improvement desired: Better bootstrap_N5 discrimination

-- What to look for: A weak funnel toward near-native models, with multiple non-native-like false positive models. This is due to the two carbohydrate units that are exocyclically connected to the glycoligand chain having multiple low-energy conformations

*5ZHO (unbound protein 5ZHG)

-- G4P RVC VP8* - A histo-blood group

-- # near-natives: 57

-- N5: 5

-- bootstrap_N5: 4.746

-- bootstrap_N5 std: 0.528

-- Future improvement desired: More near-native models

-- What to look for: A weak funnel toward near-native models, but with many near-native models generated. Potentially will see non-native-like models up to 12 Ang scored as favorably as the native-like models

*6R3M (unbound protein 1V0A)

-- CtCBM11 - beta-1,3-1,4-mixed-linked tetrasaccharide

-- # near-natives: 59

-- N5: 5

-- bootstrap_N5: 3.322

-- bootstrap_N5 std: 1.189

-- Future improvement desired: Better bootstrap_N5 discrimination

-- What to look for: A funnel toward near-native models, but with a handful of potential false positives within 2-4 Ang ring-RMSD

Target 5OYE has a bootstrap_N5 cutoff of >= 2 (instead of >= 3) in this scientific test in order to catch a single case where GlycanDock performance on more difficult targets increases, if it does.

DEFINITIONS AND COMMENTS

Nothing will work without the -include_sugars flag!

The input structures used here have already been run through Rosetta once (scored and a .pdb file dumped) so that the carbohydrate residues are converted into Rosetta formatting. This process requires flags such as -include_sugars -lock_rings -alternate_3_letter_codes pdb_sugar -maintain_links.

If the sugar has a branch point, there MUST be a TER record that separates the branched residue from the rest of the carbohydrate chain. See 2J1U and 5OYE for an example.

I used the ref2015 scorefunction for benchmarking only. Note that the GlycanDockProtocol.cc file sets fa_intra_rep_nonprotein to the weight of fa_rep (generally 0.55) and sugar_bb to 0.5 (but could be overrided if user provides different weights in the flags file).

-lock_rings if set to True by default, meaning that minimization should NOT alter carbohydrate nu angles.

Strengths of the benchmark are that (1) the protein structures used are the unbound coordinates (more realistic) and (2) the initial conformation of the glycoligand is deviated up to 7 Ang ring-RMSD (also more realistic)

LIMITATIONS

Only 6 of the original 65 targets are tested here, with a bias toward CBM/lectin protein receptors.

Notably, in the benchmark paper, each target for each initial ring-RMSD bin had 10 unique starting conformations (e.g. initial ring-RMSD bin 7 +/- 0.1 Ang target 1ABC input-1.pdb, input-2.pdb, ..., input-10.pdb). Here in this scientific test, only one input conformation is used as input for each target. The single input structure chosen per target that resulted in the best performance in the benchmark paper was chosen as the representative input structure per target here in this sci test. Therefore, the results per target are biased toward out-performing the results seen in the benchmark paper (at least in terms of # of near-native models generated).

This scientific test uses nstruct=1000 for time consideration instead of the nstruct=2000 used in the benchmark paper.

More targets from the benchmark that represent other protein receptors of interest, such as an antibody

More computing power would allow for more models generated, preferably up to nstruct=2000 to match the benchmark

Running a parallel of this benchmark using the beta_nov16 scorefunction to see performance differences