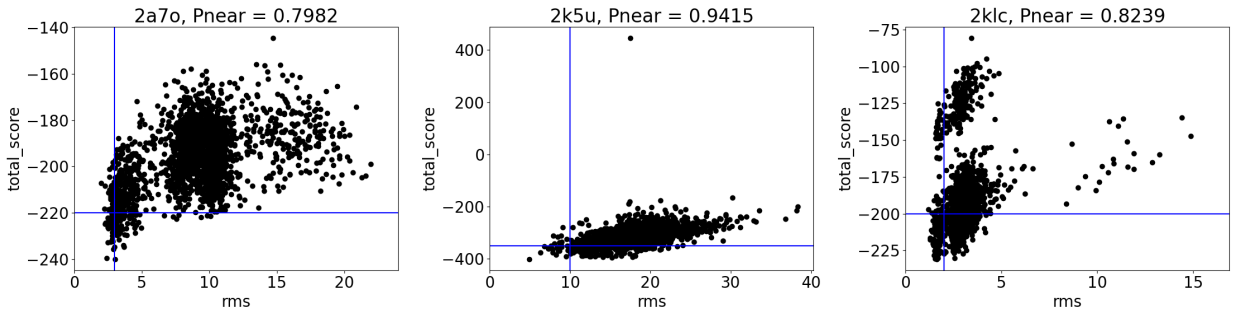
Scientific test: abinitio_RosettaNMR_rdc

FAILURES

None

RESULTS



AUTHOR AND DATE

The benchmark was originally created by Georg Kuenze (georg.kuenze@gmail.com), published in 2019, former Meiler lab, now at Leipzig University. It was implemented on the test server by Julia Koehler Leman (julia.koehler.leman@gmail.com) in the Bonneau lab, in July 2021.

PURPOSE OF THE TEST

The benchmark ensures that the score-vs-rmsd distribution of ab initio models created with NMR data (RDC / PCS specifically) don't shift too much from the original distribution.

BENCHMARK DATASET

The benchmark set contains 3 proteins of various sizes:

2klc - alpha / beta protein, 101 residues, 75 RDC constraints

2a7o - alpha protein, 112 residues, 120 RDC constraints

2k5u - alpha / beta protein, 181 residues, 322 RDC constraints

The benchmark set is described in detail in (Kuenze, Structure, 2019). The protocol runs ab initio structure prediction with NMR constraints, so input files are essentially a fasta sequence and constraint files. All input files are located in the scientific data submodule with the folder of the same name. Input files were originally taken from the protocol capture. Input files are the following:

.fasta - containing the sequence

.pdb - as a native for RMSD comparison

fragments (3mers / 9mers) - created with chemical shift information from TALOS - these are the .tab files in the scientific data submodule

.wts_patch - these are patch files containing the weight of the the RDC score against the rest of the Rosetta scorefunction terms, see below for how they are determined

.tbp - topology broker file for ab initio structure prediction

.rdc.inp - RosettaNMR constraint files containing the mathematical details of the RDC constraint setup, for instance the alignment tensor. This file contains the .dat file names that contain the actual measured RDCs.

PROTOCOL

In a nutshell, the overall protocol comprises of the following steps:

1) use chemical shift information to run TALOS for the prediction of secondary structure

2) use secondary structure prediction from TALOS files for fragment picking

3) run ab initio structure prediction WITHOUT NMR constraints to get a baseline of the score distribution

4) rescore these decoys with NMR constraint data to get score distribution WITH NMR data

5) from both score distributions of the models WITH and WITHOUT NMR data, compute the optimal weight of the NMR score term

6) run ab initio structure prediction WITH NMR constraints with the optimized weight

Note that the protocol on the test server only runs the last step (step 6) and both fragment picking and constraint weight optimization has been done beforehand.

Runtimes are about 170 CPU hours for this test: (100s per model per target) x (3 targets) x (2000 decoys)

PERFORMANCE METRICS

Output files of structure prediction are a score file and a binary silent file. We look at the score file and plot the score-vs-rmsd distribution of the created models. Passes are defined by the cutoffs for all to be true: 10% of the models below the RMSD cutoff, 10% of the scores below the score cutoff and PNear higher than the PNear cutoff, defined by the first run minus 0.1. All cutoffs were defined by the first run of the protocols and adjusted over several runs.

KEY RESULTS

We compare the results against the benchmarks described in (Kuenze, Structure, 2019). 2k5u doesn't sample as low RMSDs as in the paper, from the first, single run it is unclear why. Further investigation will be required.

DEFINITIONS AND COMMENTS

LIMITATIONS

The run times are on the higher end, which is why we're only testing 3 proteins on the test server. Ideally, it would be nice to run all benchmarks from the paper on the test server, but this is computationally prohibitive. Target diversity, size and complexity of the full benchmark are well-chosen and optimized.