Scientific test: simple_cycpep_predict

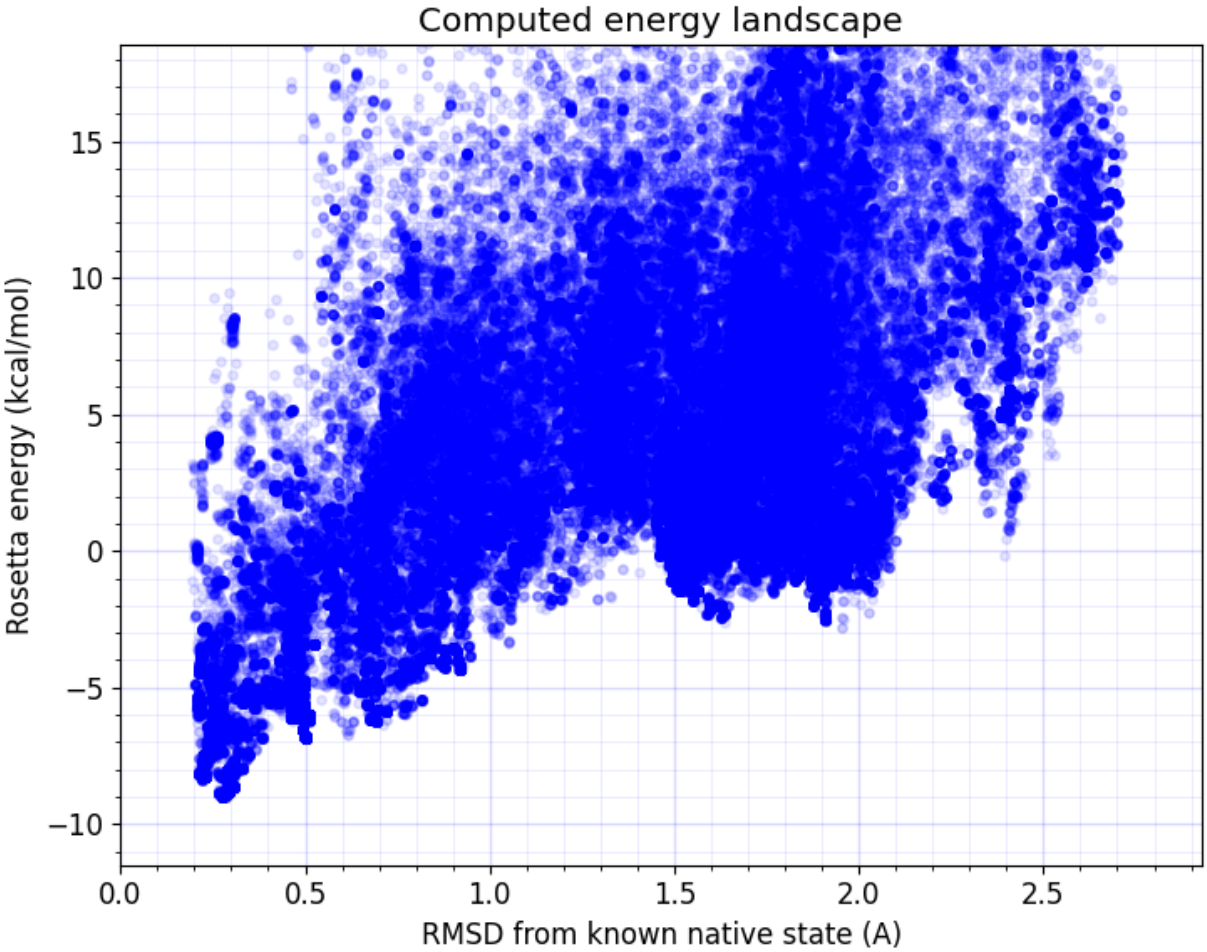
FAILURES

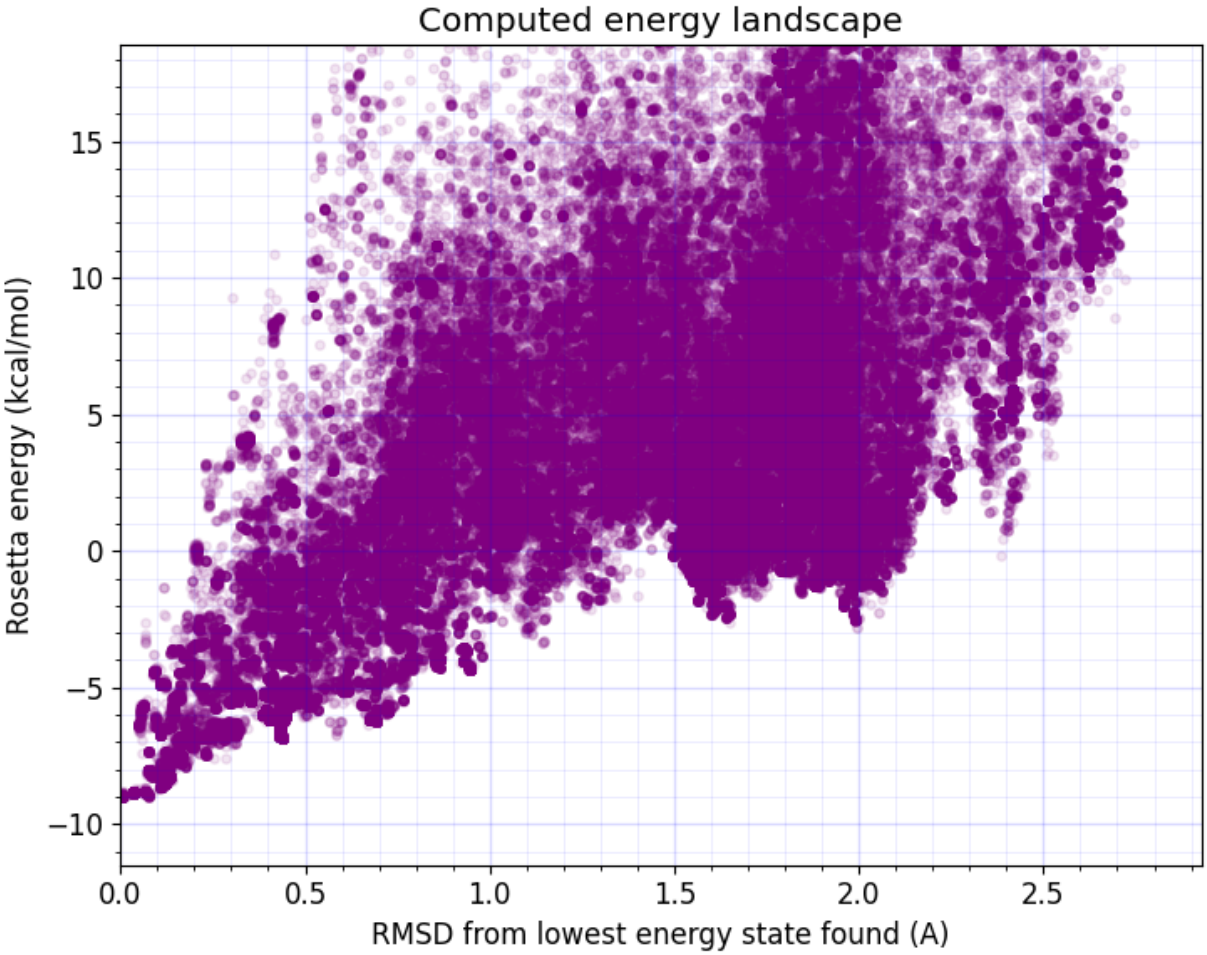
None.

RESULTS SUMMARY

Total samples = 694335
Computed PNear = 0.94544
Computed PNear to lowest E = 0.98525
Lowest energy = -8.99346 kcal/mol
RMSD of lowest energy = 0.274278 Angstroms
Lowest RMSD = 0.19352 Angstroms
Highest RMSD = 2.78773 Angstroms
Energy gap (minE>1.5A - minE) = 6.193300000000001 kcal/mol

More than 230,000 samples? YES
PNear value over 0.94? YES
PNear value to lowest E over 0.97? YES
Lowest energy under 0.3 A RMSD? YES
Sampling below 0.25 A RMSD? YES
Sampling beyond 1.5 A RMSD? YES
Sampling beyond 2.6 A RMSD? YES
6+ kcal/mol energy gap? YES
OVERALL PASS? YES


Computed energy landscape
Rosetta energy (kcal/mol)
15
10
5
0
−5
−10
0.0
0.5
1.0
1.5
2.0
2.5
RMSD from known native state (A)



AUTHOR

Vikram K. Mulligan (vmulligan@flatironinstitute.org), Center for Computational Biology, Flatiron Institute, 1 March 2019

PURPOSE OF THE TEST

This test ensures that the simple_cycpep_predict application retains its scientific performance. This application is intended to predict the structures of peptide macrocycles built from arbitrary canonical or non-canonical building-blocks.

What does the benchmark test and why?

This test case is a peptide known to fold into the designed structure, for which multiple crystal structures have been solved. The application samples peptide conformations and attempts to predict the native state from sequence alone. The test is successful if:

- The application samples close to the native state.

- The lowest-energy sample is near native.

- The funnel quality metric PNear is greater than 0.9. (This metric ranges from 0.0, for a molecule that does not favour the designed state at all, to 1.0, for a molecule that spends all of its time in the designed state. It is calculated automatically from the sampled ensemble in MPI mode by the simple_cycpep_predict job distributor.)

BENCHMARK DATASET

How many proteins are in the set?

- One peptide, nicknamed "Moriarty".

What dataset are you using? Is it published? If yes, please add a citation.

- This peptide is currently unpublished.

What are the input files? How were the they created?

- The input is the peptide, designed with Rosetta, in PDB format, along with its sequence in ASCII text format.

PROTOCOL

State and briefly describe the protocol.

The simple_cycpep_predict application uses the generalized kinematic closure algorithm (GenKIC) to rapidly sample closed conformations of a heteropolymer macrocycle built from any combination of alpha-amino acids, peptoids, or other related building-blocks. Each closure attempt is relaxed using the FastRelax protocol. For small (~8 to ~10 residue) peptide macrocycles, the application can usually sample close to the native state with less expense than protein _ab initio_.

The simple_cycpep_predict application supports hierarchical MPI-based job distribution and data reduction, as well as multi-threaded parallel job execution within a node. In MPI mode, statistics about the full sampled ensemble, including the funnel quality metric PNear, are computed automatically during the data collection and reduction phase.

Is there a publication that describes the protocol?

The simple_cycpep_predict application is described in the following publications:

1. Bhardwaj G, Mulligan VK, Bahl CD, _et al._ (2016). Accurate de novo design of hyperstable constrained peptides. _Nature_ 538(7625):329-35.

2. Hosseinzadeh P, Bhardwaj G, Mulligan VK, _et al._ (2018). Comprehensive computational design of ordered peptide macrocycles. _Science_ 358(6369): 1461-6.

How many CPU hours does this benchmark take approximately?

320 CPU-hours. The test runs on 4 nodes, 20 cores per node, for 4 wall hours. Note that job distribution ends after 2.5 hours, and results collection and analysis is expected to take less than an hour, so the actual cost is slightly lower.

In debug mode, this test takes 30 CPU-hours (1 node, 20 cores, for 1.5 wall hours). In this case, job distribution ends after 15 minutes, and results collection and analysis can take less than 75 minutes, so the actual cost is somewhat lower.

PERFORMANCE METRICS

What are the performance metrics used and why were they chosen?

All of the following must be true for the test to pass:

- More than 50,000 samples (1,500 in debug mode).

- Lowest-RMSD sample < 0.25 A from native.

- Highest-RMSD sample > 2.6 A from native.

- Lowest-energy sample < 0.3 A from native.

- Energy gap (gap between lowest-energy sample > 1.5 A and overall lowest) bigger than 6 kcal/mol

- PNear > 0.92.

How do you define a pass/fail for this test?

Failure of any of the above.

How were any cutoffs defined?

Arbitrarily, like so much else in Rosetta. These are based on the performance of Rosetta in predicting the crystal structure

of this peptide on 27 June 2019.

KEY RESULTS

What is the baseline to compare things to - experimental data or a previous Rosetta protocol?

Past iterations of this test.

Describe outliers in the dataset.

N/A.

DEFINITIONS AND COMMENTS

State anything you think is important for someone else to replicate your results.

N/A.

LIMITATIONS

What are the limitations of the benchmark? Consider dataset, quality measures, protocol etc.

We have very few known crystal structures of cyclic peptides.

How could the benchmark be improved?

More peptides. (We will add more in the future.)

What goals should be hit to make this a "good" benchmark?