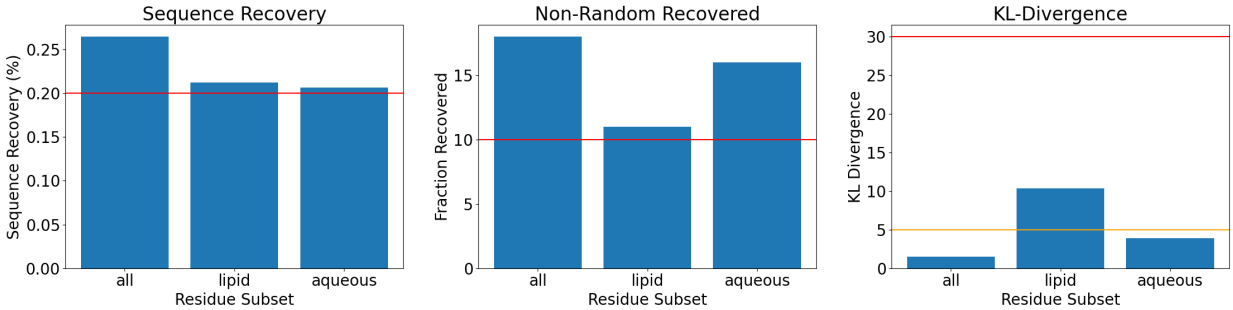
Scientific test: mp_f19_sequence_recovery

FAILURES

None

RESULTS



AUTHOR AND DATE

Rebecca F. Alford (ralford3@jhu.edu)

PI: Jeffrey J. Gray (Johns Hopkins ChemBE)

Test created 6/6/19

PURPOSE OF THE TEST

The purpose of this test is to evaluate the scientific performance of franklin2019, the default energy function for membrane protein structure prediction and design.

BENCHMARK DATASET

The benchmark dataset includes 130 alpha-helical and beta-barrel transmembrane proteins, with <25% sequence identity and better than 3.0Ã… resolution. The dataset is a subset of proteins from [1] which have assigned lipid compositions. The dataset modifications are detailed in [2].

([1] Koehler Leman J, Lyskov S, Bonneau R (2017) "Computing structure-based lipid accessibility of membrane proteins with mp_lipid_acc in RosettaMP" BMC Bioinformatics 18:115)

([2] Alford, R. F., Fleming, P. J., Fleming, K. G. & Gray, J. J. Protein Structure Prediction and Design in a Biologically Realistic Implicit Membrane. Biophys. J. 118, 2042â€“2055 (2020).)

The inputs are PDB coordinate files and spanning topology definition files for each protein. The PDB coordinate input files were downloaded from the

Orientations of Proteins in Membranes Database. The spanning topology definition files were generated using the mp_span_from_pdb application.

PROTOCOL

To evaluate sequence recovery, the fixed-backbone Rosetta design protocol is used to search for low energy sequences. The protocol is described in:

(Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, Thompson J, Davis IW, Pache RA, Lyskov S, Gray JJ, Kortemme T, Richardson JS, Havranek JJ, Snoeyink J, Baker D, Kuhlman B (2013) "Scientific benchmarks for guiding macromolecular energy function improvement" Methods in enzymology 523:109-143)

The benchmark will take approximately 500 CPU hours.

PERFORMANCE METRICS

To evaluate the performance of this benchmark, we computed three metrics. The first metric is sequence recovery which is the fraction of correctly designed positions relative to the number of available positions. Realistic energy functions will maximize the recovery rate, with ideal values ranging from 35-50%. Note, perfect sequence recovery is seldom possible because other factors constrain protein sequences including functional and evolutionary pressures.

The second metric is the recovery rate for individual amino acids relative to the background probability of guessing a random amino acid type (1 in 20 types, or 5%). Here, a higher value is better.

The third metric is the Kullback-Leibler divergence which is a measure of the divergence of the amino acid distribution in the designed sequences from the distribution in the native sequences. Unlike recovery and non-random rates, the goal is to minimize the KL-divergence. Ideal values for a membrane protein set are under 5.0, delineated by the yellow solid line on the plot.

For sequence recovery, pass/fail is defined by comparing newly computed with established values computed in [Alford et al. 2020: Protein structure prediction...], which 0.2. For Non-random recovery, a passing value is greater than 10%. A KL divergence failure is defined by a value < 5.0 for the 'all' subset and if both subsets lipid and aqueous have a KL divergence < 5.0

KEY RESULTS

The key results of this scientific test are twofold:

(1) Sequence recovery is high for all amino acid types, not just non-polar amino acids in the transmembrane region

(2) The fraction of amino acid types recovered with rates higher than random, Naa. Naa is > 75%, compared with older energy functions mpframework_fa_2007 (Barth et al. 2007) and

mpframework_smooth_fa_2012 (Yarov-Yaravoy et al. 2006) for which Naa was generally less than 50%. This previously resulted in design skewed toward nonpolar amino acids, rather than sampling from a diverse palette of chemistries.

DEFINITIONS AND COMMENTS

The transmembrane, interface, and bulk solvent, as well as buried vs. surface exposed criteria are described in the following paper:

(Alford, R. F., Fleming, P. J., Fleming, K. G. & Gray, J. J. Protein Structure Prediction and Design in a Biologically Realistic Implicit Membrane. Biophys. J. 118, 2042â€“2055 (2020).)

LIMITATIONS

It would be great to analyze the data without a Rosetta executable, which takes longer to debug.

In general, the benchmark should be more balanced between alpha-helical and beta barrel membrane proteins. The dataset is currently ~25% beta-barrel proteins.