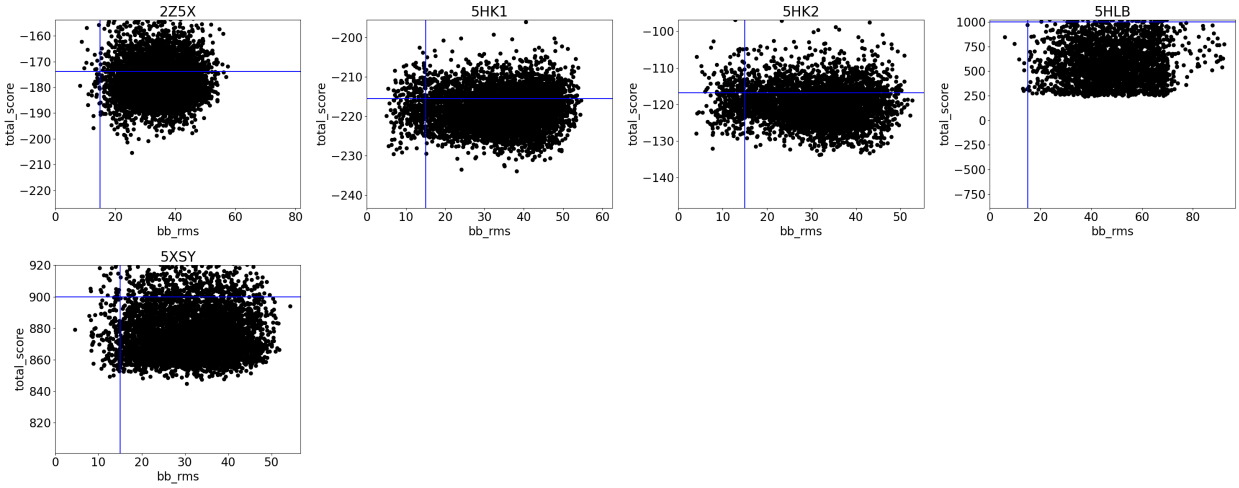
Scientific test: mp_domain_assembly

FAILURES

None

RESULTS



AUTHOR AND DATE

The benchmark was set up by Julia Koehler Leman (julia.koehler.leman@gmail.com) in March 2019.

The PI is Richard Bonneau.

PURPOSE OF THE TEST

This is the first method for domain assembly for membrane proteins. You can give it structures or models of protein domains and a fasta sequence and it assembles the protein into a full-length model. There aren't a lot of structures of full-length membrane proteins, this is why the benchmark set only consists of 5 proteins so far. Also, since the domains are mostly connected by flexible linkers, the energy landscape is pretty flat along those linkers, so the only thing we can test so far is how well we sample. We basically only check whether we can sample a model <10A RMSD. Also, the RMSD values get pretty large pretty quickly, because only the TM domains are superimposed, which in this benchmark set are single TM helices, which make up a minority of the protein in terms of number of atoms.

BENCHMARK DATASET

The benchmark set consists of 5 proteins - the benchmark set, method and command lines are published in (Koehler Leman & Bonneau, Biochemistry, 2017).

Structures were downloaded from the PDBTM database, where the proteins are transformed into the membrane coordinate system. We included only structures without gaps as the fasta files are created from the ATOM lines in the PDB. Structures were cleaned, the spanfile was created with mp_span_from_pdb, the membrane embedding was optimized with mp_transform, a fasta file was created from the ATOM lines, and the fasta file was used to pick fragments using Robetta, excluding homologues. These structures are the natives we compare our models to.

We then removed a few residue linkers and split the PDB files into TM domain and soluble domain. This was done via visual inspection. Linker lengths are:

2z5x: 10 res

5hk1: 6 res

5hk2: 6 res

5hlb: 9 res

5xsy: 9 res

These are used as input files. Input files for the protocol are:

-in:file:fasta 2Z5X_tr_A.fasta # fasta file

-in:file:frag3 2Z5X.frags.3.200_v1_3 # 3-residue fragments

-in:file:frag9 2Z5X.frags.9.200_v1_3 # 9-residue fragments

-in:file:native 2Z5X_tr_A_opt.pdb # native for RMSD calculation (this is done without superposition, i.e. only the TM domain remains superimposed)

-mp:setup:spanfiles 2Z5X_tr_A.span # required for RMSD calculation with native

-mp:assembly:poses 2Z5X_tr_A_opt_sol.pdb 2Z5X_tr_A_opt_tm.pdb # structures of the input domains

PROTOCOL

The protocol is described in detail in (Koehler Leman & Bonneau, Biochemistry, 2017). Briefly, it starts with the TM domain embedded in the membrane, linker residues are added towards the N-terminus, then the N-terminal domain is added, then linker residues towards the C-terminus is added, then the C-terminal domain is added to create the full-length pose. Note that this benchmark set contains only structures with one soluble domain but the protocol works with soluble domains at either terminus (it should also work with a more domains on either side, like beads on a string, the TM domain somewhere in the middle). There is an optional refinement step that runs after domain assembly, but this step was forgone for this benchmark set because of runtime constraints. We create 5000 models for each protein.

For this benchmark set, a model is generated in ~100s. 100s x 5 proteins x 5000 models = 2.5 megaseconds is <700 CPU hours.

PERFORMANCE METRICS

Since the energy landscape of flexible linkers is quite flat, it is difficult to identify a near-native model by score alone. Relaxed natives have scored better than built models, as shown in the paper, figure 2, but sampling so close to the native remains a challenge. We therefore check for RMSDs<15A for at least 2 of the models, which is 0.06%. For score, we ask whether 3 of the models are below the score cutoff which was set from the avg score during the first run, then adjusted visually. RMSD values remain large due to the lever arm effect. Cutoffs were defined by running the protocol a couple of times and adjusting the cutoffs for the test to pass.

KEY RESULTS

The proteins are all single TM helix proteins with a large soluble domain - we compare to crystal structures from the PDB. 5HK1 and 5HK2 are basically the same protein with slightly different relative domain orientations. Please also compare to figure 2 in (Koehler Leman && Bonneau, Biochemistry, 2017).

Since the energy landscape is relatively flat, we don't see a pronounced funnel, unless you include relaxed natives into the plot.

DEFINITIONS AND COMMENTS

The protocol is currently being improved for modeling dimers and all kinds of combinations of soluble and TM domains; it also includes better error handling for identifying gaps and mutations.

LIMITATIONS

It remains challenging to sample sufficiently close to the native structure, even with 100k models. The benchmark set is currently rather small. Once more structures become available, it should be updated. Quality measures could also be improved. RMSD is not a good measure as it becomes large quickly due to the lever arm effect. However, GDT might not be a good measure either since have the structures of the individual domains, which would result in a large GDT to begin with, and all we are interested in is the correct modeling of the linkers. Maybe a dihedral angle distribution might be the way to go. Ideas are appreciated, please email Julia.