Distill, Distill_roll

Distill for CASP10

C. Mirabello, G. Tradigo, P. Veltri, and G. Pollastri

1 – UCD Dublin, Ireland, 2 – Università di Cosenza, Italy

gianluca.pollastri@ucd.ie

Distill has two main components: a fold recognition stage dependent on sets of protein features predicted by machine learning techniques; an optimisation algorithm that searches the space of protein backbones under the guidance of a potential based on templates found in the first stage. The main differences with our CASP9 systems are: the greatly improved fold recognition stage; the fact that we fit structures directly to the distance maps of templates rather than to predicted contact maps. The difference between Distill and Distill_roll is that for the former we use an improved fold recognition algorithm.

Methods
Distill runs 3 rounds of PSI-BLAST against a 90% redundancy reduced UniProt to generate multiple sequence alignments (MSA). The PSSM from the second round is reloaded to search the PDB for templates (e=1e-3). MSA and templates are fed to our 1D prediction systems (all based on BRNN): Porter1,4 (secondary structure), PaleAle4 (solvent accessibility), BrownAle4 (contact density), Porter+ (structural motifs). All predictors use template information as an input alongside the sequence and MSA.

1D predictions are combined into a structural fingerprint (SAMD) which, alongside the PSSM, is used to find remote homologues in the PDB through 3 searches for Distill_roll (PSSM and SAMD profile against PDB sequences and SAMD, with 3 different substitution matrices) and 6 searches for Distill (same as above, plus 3 more searches against PDB PSSM rather than sequences). In the following stage residue contact maps are predicted by a system based on 2D-Recursive Neural Networks (XXstout). We predict binary maps with a contact threshold of 8Å between Cβ, which are submitted to the RR category. Inputs for map prediction are: the sequence; MSA; PSI-BLAST, SAMD and SAMD templates. That is, the maps are template-based whenever suitable templates are found. The 3D reconstruction, which is only conducted on Cα traces, is run as follows: we run a SAMD search for templates with an e-value of 10,000; for each (overlapping) 9-mer of the protein we gather the structures of the top 50 templates which fully cover it (SAMD_list); a simulated annealing search is run using crankshaft moves to quickly find a minimum of a potential function which rewards formation of contacts that appear in a weighed average of the distance maps of templates; from the previous enpoint a simulated annealing search is run by substituting 9-mers from the conformation with 9-mers from the SAMD_list, and using the same potential function as above.

We run 30 reconstructions for each protein, which we rank by their weighed TM-scores against the template list. For the 5 top-ranked models we reconstruct the backbone with SABBAC, and the full atoms with Scwrl4, then run a brief energy minimisation by gromacs. These are the models submitted to CASP.

Results
We await the CASP assessment. On preliminary tests (on the CASP9 set) we have observed a GDT_TS improvement of over 5% over our CASP9 systems.

Availability
http://distillf.ucd.ie/distill/ (Distill), http://distillf.ucd.ie/distill_roll/ (Distill_roll)


Abstracts - 1