

Distill

Distill for CASP12

B.Alshomrani, M.Torrisi, M.Kaleel, G. Pollastri

UCD Dublin, Ireland

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. Apart from updating the underlying databases, Distill is unchanged since 2012.

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): Porter^{1,4,6} (secondary structure), PaleAle^{4,6} (solvent accessibility), BrownAle⁴ (contact density), Porter+² (structural motifs). All predictors use template information as an input alongside the sequence and MSA.

1D predictions are combined into a structural fingerprint⁴ (SAMD) which, alongside the PSSM, is used to find remote homologues in the PDB through 6 searches (PSSM and SAMD profile against PDB sequences and SAMD, with 3 different substitution matrices, 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 (XXstout⁵). We predict binary maps with a contact threshold of 8Å between C β , 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 C α 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 of the conformational space is run by substituting snippets of 3 to 9 amino acids extracted from the SAMD_list 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 low temperature refinement 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.

It should be noted that everything in our pipeline (except BLAST and the software to blow C α traces into full-atom models) is in house, and that in normal conditions we can provide predictions for a protein in tens of minutes.

Results

We await the CASP assessment.

Availability

<http://distillf.ucd.ie/distill/>

1. Pollastri, G. & McLysaght, A. (2005) Porter, A new, accurate server for protein secondary structure prediction, *Bioinformatics*, **21**(8), 1719–1720.
2. Mooney, C., Vullo, A. & Pollastri, G. (2006) Protein Structural Motif Prediction in Multidimensional ϕ - ψ Space leads to improved Secondary Structure Prediction, *Journal of Computational Biology*, **13**(8), 1489-1502.
3. Walsh, I., Martin, A.J.M., Mooney, C., Rubagotti, E., Vullo, A. & Pollastri, G. (2009). Ab initio and homology based prediction of protein domains by recursive neural networks" *BMC Bioinformatics*, **10**, 195.
4. Mooney, C. & Pollastri, G. (2009). Beyond the Twilight Zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information, *Proteins*, **77**(1), 181-90.
5. Walsh, I., Baú, D., Martin, A.J.M., Mooney, C., Vullo, A. & Pollastri, G. (2009). Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks, *BMC Structural Biology*, **9**, 5.
6. Mirabello, C. & Pollastri, G. "Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility", *Bioinformatics*, 29(16):2056-2058, 2013