Distill for CASP13

M.Torrisi[1], M.Kaleel[1] and G.Pollastri[1]

[1] - *School of Computer Science, University College 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. The residue contact maps submitted by Distill are predicted fully ab initio by an ensemble of 2D-Recursive Neural Networks trained on evolutionary features including correlated mutations.

Methods

Distill runs PSI-BLAST and hhblits against recent redundancy reduced versions of UniProtKB to generate multiple sequence alignments (MSA). The PSSM from the PSI-BLAST search is reloaded to search the PDB with PSI-BLAST for an initial guess at templates. MSA and templates are fed to our 1D prediction systems (all based on stacks of Bidirectional Recurrent Neural Networks and Convolutional Neural Networks): Porter[1,4,6,7] (secondary structure), PaleAle[4,6] (solvent accessibility), BrownAle[4] (contact density), Porter+[2] (structural motifs). All predictors use template information as an input alongside the sequence and MSA. The ab initio components of all predictors have recently been trained anew on sets of roughly 15,000 protein structures extracted from the PDB and should be considerably improved compared with the versions adopted at previous CASP editions.

1D predictions are combined into a structural fingerprint[4] (SAMD) which, alongside the PSSM, is used to find remote homologues in the PDB through 6 Smith-Waterman searches (PSSM and SAMD profile against PDB sequences and SAMD, with 3 different substitution matrices, plus 3 more searches against PDB PSSMs rather than sequences).

In parallel, residue contact maps with a contact threshold of 8Å are predicted by a newly trained system based on 2D-Recursive Neural Networks[5], and submitted to the RR category. Inputs for map prediction are: profiles from MSA; outputs from freecontact, CCMpred; selected 1D and 2D statistics from the MSA used. That is, the maps are always purely ab initio unlike Distill versions for previous CASP editions.

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 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 agreement with a set of desired constraints for the protein (see below); from the previous endpoint 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. The set of desired constraints driving the protein reconstruction is a weighted average of the distance maps of templates, interpolated, where templates are missing, with predicted ab initio maps as submitted to the RR category. That is, if no templates are found the reconstruction is purely based on our predicted contact map.

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

Results

We await the CASP13 assessment.

Availability

The newest version of Distill is available at http://distilldeep.ucd.ie/casp/

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. (2013) Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility, *Bioinformatics*, **29**(16):2056-2058.
7. Torrisi, M, Kaleel, M & Pollastri, G. (2018) Porter 5: fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes, *bioRxiv* 289033; doi: https://doi.org/10.1101/289033