Distill, Distill_human

Distill: protein structure prediction by Machine Learning

C. Mirabello[1], G. Tradigo[1,2] and G. Pollastri[2]
[1] – UCD Dublin, Ireland, [2] – Università della Magna Græcia, Italy
gianluca.pollastri@ucd.ie

Distill has two main components: a set of predictors of protein features based on machine learning techniques; an optimisation algorithm that searches the space of protein backbones under the guidance of a potential based on these features. For CASP9 we have retrained and updated our prediction methods and fold recognition module, and our optimisation algorithm for the conformational search, which now uses "snippets" of PDB structures suggested by our fold recognition algorithm.

The only difference between Distill and Distill_human is that for the latter we evaluated and partially re-ranked Distill's models visually.

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] (secondary structure), PaleAle[4] (solvent accessibility), BrownAle[4] (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 fingerprint[4] (SAMD) which, alongside the PSSM, is used to find remote homologues in the PDB (1-against-all alignment). If this search returns templates that are deemed to be more reliable than the PSI-BLAST ones, all 1D predictions are run again with the new templates as inputs.

In the following stage residue distance and contact maps are predicted by a system based on 2D-Recursive Neural Networks (XXstout[5]). Two types of maps are predicted: binary maps with a contact threshold of 8Å between Cβ, which are submitted to the RR category; 4-class distance maps (thresholds of 8, 13 and 19Å) between Cα which are used for 3D prediction. Inputs for map prediction are: the sequence; MSA; PSI-BLAST 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 of the conformational space is run using crankshaft moves to quickly find a minimum of a potential function which rewards formation of predicted contacts; 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 templates. For the 5 top-ranked models we reconstruct the backbone with Maxsprout (with SABBAC for Distill_human), and the full atoms with Scwrl4. These are the models submitted to CASP.

It should be noted that everything in our pipeline (except BLAST and the software to blow Cα 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

For many proteins, our results seem to be competitive to us based on the first 80 structures released, but we await the CASP assessment for this.

Availability

http://dbstill.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.