A7D

De novo structure prediction with deep-learning based scoring

R.Evans[*,1], J.Jumper[*,1], J.Kirkpatrick[*,1], L.Sifre[*,1],
T.F.G.Green[1], C.Qin[,1], A.Zidek[1], A.Nelson[1], A.Bridgland[1],
H.Penedones[1], S.Petersen[1], K.Simonyan[1], S.Crossan[1], D.T.Jones[2],
D.Silver[1], K.Kavukcuoglu[1], D.Hassabis[1], A.W.Senior[*,1]

* - *Equal contribution, 1- DeepMind, London, UK; 2 UCL, London, UK.*
andrewsenior@google.com

A7D CASP13 submissions were produced by three variants of an automatic free-modelling structure prediction system relying on scores computed with deep neural networks. Scoring relied on one of two neural networks: a predictor of inter-residue distances and a direct-scoring network. The basic method used a generative neural network for fragment generation for fragment assembly in memory-augmented simulated annealing. Successive rounds of simulated annealing used fragments from the memory. The third method used full-chain score minimization with gradient descent.

Methods

The systems tested all use multiple sequence alignments (MSA) and profiles generated from HHBlits [2] and PSI-BLAST [3]. No templates were used, nor were server predictions. No manual intervention was made except for domain segmentation of T0999 and final decoy ranking in a handful of cases. In protein complexes, each chain was processed independently.

Scoring

Two neural networks were used for scoring. For the first, a very deep residual convolutional neural network was trained on a non-redundant database of proteins selected from the Protein Data Bank (PDB) to predict the distances between C-beta atoms of different residues, using MSA-based features. With these predictions and a reference distribution, a likelihood score was computed for candidate structures according to the realised distances.

A second deep residual convolutional neural network was trained to directly output a score as a function of structure geometry, MSA-based features and the contact predictions from the first network.

Domain segmentation

Domain segmentation hypotheses for two or three domains were generated by automatic analysis of the full-chain contact matrix prediction derived from the inter-residue distance prediction. Each domain segmentation hypothesis (as well as full chain without segmentation) was folded independently up to eight times with the domains in each hypothesis being folded independently.

Fragment assembly

Two approaches were used for structure modelling. The first was based on fragment assembly. For each domain, a DRAW [4] model of backbone torsion angles, trained on the same PDB subset was sampled to generate a set of overlapping 9-residue fragments. Fragments were inserted with simulated annealing using a score based on our distance predictions for the domain hypothesis plus Rosetta's [1] score2 (Variant 1) or the direct structure scoring without Rosetta (Variant 2).

Repeated rounds of simulated annealing were run, using evolutionary hyper-parameter optimization to tune run-length and start temperature, with successive rounds using fragments from the structures generated in previous rounds.

The best-scoring structures from simulated annealing were relaxed using Rosetta fast relax with our inter-residue distance prediction score and Rosetta's full-atom score.

Domain assembly

After domain-level relaxation, for each domain segmentation, full-chain structures were assembled from domain structures with simulated annealing and further relaxed. The best-scoring full-chain structure for each run of each domain segmentation hypothesis for each method was chosen.

Direct structure optimization

An alternative structure modelling approach was used for Variant 3 without any domain segmentation. Here we used gradient descent of a combination score (inter-residue distance prediction score + neural-network-based torsion angle prediction likelihood + score2) to optimize *full chain* structures, parameterised with torsion angles.

Decoy selection

Five ranked structure predictions were submitted for each "all groups" target. Initial submissions used variants 1 & 2 in parallel, but submissions from T0975 on used variants 1 & 3 in parallel. The 5 candidate submissions were the best scoring from among the independent runs of the two different methods, with a bias towards selecting from variants 2 or 3, and manual ranking in a handful of cases.

Acknowledgements

We are grateful for fruitful discussions with Oriol Vinyals and contributions from Marek Barwinski, Ruoxi Sun, Carl Elkin, Peter Dolan, Matthew Lai and Yujia Li as well as many others at DeepMind.

We also gratefully acknowledge the use of several tools and datasets (Rosetta, HHBlits, HHPred, PSI-BLAST, PDB, CATH).

1. Das,R., Baker,D. (2008) Macromolecular modeling with rosetta. *Annu Rev Biochem.* 77:363–382.
2. Remmert,M., Biegert,A., Hauser,A. (2012) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment *Nature Methods*, 9(2), 173-175.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
4. Gregor,K., Danihelka,I., Graves,A., Rezende,D.J. (2015) DRAW: A recurrent neural network for image generation *arXiv*:1502.04623