Ranking C alpha traces with Neural Network Pairwise Interaction Fields

Martin AJM[1], Pollastri G[2], Tosatto S[1]

Motivation

In order to use a 3D model of a protein structure we need to know how good it is, as its quality is proportional to its utility [1]. Several different potential or (pseudo-)energy function have been developed aiming to predict model quality. We present here an update of a knowledge-based ModelQuality Assessment Program (MQAP) at the residue level which evaluates single protein structure models [2]. We use a tree representation of the C alpha trace to train a novel Neural Network Pairwise Interaction Field (NN-PIF) to predict the global quality of a model. All the inputs to NN-PIF are derived from the C alpha trace of the models and the sequence of amino Acids associated to it.

Methods

Protein model quality is often measured as the scaled distance between C alphas of models to their positions in the native structure after optimal superimposition of the structures. Here only information obtained solely from the C alpha trace is used. First, the C alpha trace of each structure model is represented as a directed acyclic graph (rooted tree), in which the outer nodes are pairwise interactions. Each residue in the C alpha trace is encoded into a vector describing its environment. Interactions among C alphas are simply characterised by distances and angles, alongside the two vectors encoding the residues involved. Environments are described by several angles, distances among neighbours, pseudo-Solvent Accessibility (SA), and coarse packing information. All these numerical descriptors computed from the C alpha trace are fed into NN-PIF trained to predict global quality. In NN-PIF each C alpha (i.e. its interactions with all the other C alphas) is mapped into a hidden state, which contains the contribution of that residue to the global quality of the structure. Two C alphas are considered as interacting if they are closer than a fixed distance threshold (here it used 20A. The hidden vectors for all C alphas are then combined and mapped to a global quality measure. NN-PIF allows us to evaluate all the interactions at the same time, whereas other knowledge based potentials generally evaluate interactions separately. To train the NN-PIF models submitted to previous CASP editions [3] are used, as the main purpose of this MQAP is to rank models from dierent prediction systems. No native structures are included in the training set. Tests are

[1] Dipartimento di Biologia, Università degli Studi di Padova Viale G. Colombo 3, 35131 Padova [2] University College Dublin, Complex and Adaptive Systems Laboratory, Belfield, Dublin 4 Ireland

performed on CASP8 server models, a subset of the PDB REDO [4] database with significantly different C traces to their PDB [5] counterparts and several standard decoys datasets available at the Decoys'R'Us repository[6].

Results

In our tests on a large set of structures, our model outperforms most other single model evaluation methods based on different and more complex protein structure representations in both local and global quality prediction in a real scenario simulation. NN-PIF is also tested on its ability to select identify better native structures and native structures among artificial decoys. NN-PIF shows a method dependency accuracy but identify positively better native structures as their quality increases. NN-PIF allows fast evaluation of multiple di erent C alpha trace structure models for a single protein sequence. The method is available upon request from the authors. 3D structure prediction method-specific rankers may also built by the authors upon request. NN-PIF will be soon available as a web server.

Contact e-mail

albertoj@bio.unipd.it

Supplementary information

[1] D Cozzetto, A Kryshtafovych, M Ceriani, and A Tramontano. Assessment of predictions in the model quality assessment category. PROTEINS: Struc ture, Function, and Bioinformatics, 69(Suppl 8):175-183, 2007. [2] AJM Martin, A Vullo, and G Pollastri. Neural Network Pairwise Interaction Fields for Protein Model Quality Assessment. In Learning and Intelligent Optimization, Third International Conference, volume 5851 of Lecture Notes in Computer Science, pages 235-248. Springer Berlin / Heidelberg, 2009. [3] JN Battey, J Kopp, L Bordoli, RJ Read, ND Clarke, and T Schwede. Automated server predictions in CASP7. PROTEINS: Structure, Function, and Bioinformatics, 69(Suppl 8):68-82, 2007. [4] RP Joosten, J Salzemann, V Bloch, H Stockinger, A Berglund, C Blanchet, E Bongcam-Rudlo, C Combet, ALD Costa, G Deleage, M Diarena, R Fabbretti, G Fettahi, V Flegel, A Gisel, V Kasam, T Kervinen, E Korpelainen, K Mattila, M Pagni, M Reichstadt, V Breton, IJ Ticklei, and G Vriend. PDB REDO: automated re-refinement of X-ray structure models in the PDB. Journal of Applied Crystallography, 42:376{384, 2009. [5] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The protein data bank. Nucleic Acids Research, 28 (1) :235-242, 2000. [6] M Levitt and R Samudrala. Decoys'R'Us: A database of incorrect protein conformations to improve protein structure prediction. Protein Science, 30(300):1399-1401, 2000.