Prediction and cleaning of ten residue-residue distance thresholds using machine learning

Walsh I[1], Tosatto S[1], Pollastri G[2]

Motivation

Prediction of protein 3D structure from the primary sequence remains a fundamental and extraordinarily challenging problem [1]. A contact map is a two-dimensional (2D) projection of the 3D protein. An obvious 2D projection of the 3D structure is the matrix of contacting residue pairs, or contact map. Contact maps, or similar distance restraints have been proposed as intermediate steps between the primary sequence and the 3D structure (e.g. in [2, 3, 4]), for various reasons: unlike 3D coordinates, they are invariant to rotations and translations, hence less challenging to predict by machine learning systems [4]; quick, effective algorithms exist to derive 3D structures from them, for instance stochastic optimisation methods [5, 6], distance geometry [7], or algorithms derived from the NMR literature and elsewhere [8]. Most of the literature and the Critical Assessment of protein Structure Prediction (CASP) experiments deal with binary contact maps with residue pairs (h,k) either in contact (at a certain distance threshold) or not. Previously we have shown that 4-class distance maps are more useful for reconstructing 3D models than binary contact maps [14]. Here we construct a more difficult problem 10-class maps. The experiment is a preliminary work in order to gauge if our machine learning method is capable of learning this difficult classification scheme. Filtering or cleaning predictions made about protein features have been shown to be useful for secondary structure [9,10] and contact density [11]. In [12] physical rules were manually constructed in order to clean the contact map. In this work we will also describe a machine learning method for cleaning distance maps.

Methods

2D-Recursive Neural Networks (2D-RNN) were previously described in detail in [4] and [13] where they were generally described as Directed Acyclic Graph RNN's. This is a family of adaptive models for mapping 2D matrices of variable size into matrices of the same size. Two distinct 2D-RNN's were used in order to make the final prediction. The first 2D-RNN processes the input and produces probabilities, O (h,k), about the classification. The second filter 2D-RNN produces classifications as a function of global inputs and the predicted probabilities, O (h,k), coming from

[1] Dipartimento di Biologia, Università degli Studi di Padova Viale G. Colombo 3, 35131 Padova, Italy. [2] School of computer science and informatics, Complex and Adaptive Systems Laboratory, Dublin 4, Ireland

the first 2D-RNN. The information supplied is: • The average probability of each class located in non-overlapping square windows. The first windows centre is located at the residue pair (h,k). I choose the size of the windows to be 11 and the number of non-overlapping windows considered around (h,k) to be 15. • Each class probability. • The residue wise contact order (RWCO) [15]. • The number of distance class types for residues h or k • The number of distance class types in common with both residues j and k Preliminary tests were carried out on a training set of 200 and a testing set of 60 proteins. All proteins have length less than or equal to 200 residues. The 10 class distance thresholds are [0,2), [2,4), [4,6), [6,8), [8,10), [10,12), [12,14), [14,16), [16,18), [18,infinity) angstrom. Homology information is supplied in a similar manner to [14].

Results

Accuracies and the size of each class on the test set for the 10 classes are: Accuracy Number of residue pairs [0,2) 91.0 5733 [2,4) 95.6 11470 [4,6) 73.1 17046 [6,8) 60.0 20174 [8,10) 60.1 32814 [10,12) 55.1 43566 [12,14) 48.6 50132 [14,16) 46.4 55504 [16,18) 40.1 53526 [18,inf] 93.2 323364 Total 74.6 613329 The accuracies show that our method is capable of learning this difficult problem. In addition, the filter improves the total accuracy by 0.4% (i.e. 24533 residue pairs). This work was a preliminary investigation into a difficult classification scheme. In the future we will: • Carry out an analysis on a small set (e.g. 50 proteins) to find the optimal thresholds for the 10 classes which will maximise the quality of the final 3D models. The small improvements for the filter are encouraging but further improvements are necessary in order to make the cleaning algorithm worthwhile in a sequence to 3D modelling pipeline. Future directions will include: • A more extensive investigation to find additional parameters which will improve the accuracy of the maps. • Rules such as beta-strand residues must have no more than 2 close partners could also be easily implemented. • A third machine learning layer which would maximise TM-score or minimise RMSD will also be investigated. Although, this would be computationally expensive because each predicted contact map would require a 3D model to be reconstructed. It may be a worthwhile experiment since the ultimate goal of all protein prediction methods is the final quality of the 3D model.

Contact e-mail

ian.walsh@ucd.ie

Supplementary information

References

[1] D. Baker and A. Sali. Protein structure prediction and structural genomics. Science, 294:93–96, 2001. [2] P. Fariselli and R. Casadio. A neural network based predictor of residue contacts in proteins. Protein Engineering, 12(1):15–21, 1999. [3] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Prediction of contact maps

with neural networks and correlated mutations. Protein Engineering, 14(11):835–439, 2001. [4] G. Pollastri and P. Baldi. Prediction of contact maps by recurrent neural network architectures and hidden context propagation from all four cardinal corners. Bioinformatics, 18, Suppl.1:S62–S70, 2002. [5] M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. Folding and Design, 2:295–306, 1997. [6] D.A. Debe, M.J. Carlson, J. Sadanobu, S.I. Chan, and W.A. Goddard. Protein fold determination from sparse distance restraints: the restrained generic protein direct monte carlo method. J. Phys. Chem., 103:3001–3008, 1999. [7] A. Aszodi, M.J. Gradwell, and W.R. Taylor. Global fold determination from a small number of distance restraints. J. Mol. Biol., 251:308–326, 1995. [8] J. Skolnick, A. Kolinski, and A.R. Ortiz. Monsster: a method for folding globular proteins with a small number of distance restraints. J. Mol. Biol., 265:217–241, 1997. [9] G. Pollastri and A. McLysaght. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics, 21(8):1719–20, 2005. [10] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol., 232:584–599, 1997. [11] A. Vullo, I. Walsh, G. Pollastri. A Two-stage Approach for Improved Prediction of Residue Contact Maps. BMC Bioinformatics, 7:180, 2006. [12] Y. Shao and C. Bystroff. Predicting interresidue contacts using templates and pathways. Proteins, 53:487–502, 2003. [13] P. Baldi and G. Pollastri. The principled design of large-scale recursive neural network architectures – dag-rnns and the protein structure prediction problem. Journal of Machine Learning Research, 4(Sep):575–602, 2003. [14] I. Walsh, D. Bau, A. J. M. Martin, C. Mooney, A. Vullo, G. Pollastri. Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC Structural Biology, 9:5, 2009. [15] A. R. Kinjo and K. Nishikawa. Predicting secondary structures, contact numbers, and residuewise contact orders of native protein structures from amino acid sequences using critical random networks. BIOPHYSICS Vol. 1 (2005) pp.67