Poster I-73

A Filtering Approach for Improved Modelling of Predicted Contact Maps Alberto J. Martin, Dr Alessandro Vullo, Dr Gianluca Pollastri School of Computer Science and Informatics University College Dublin Belfield, Dublin 4, Ireland {albertoj,alessandro.vullo,gianluca.pollastri}@ucd.ie

Authors:
Alberto J. Martin (*School of Computer Science and Informatics University College Dublin*)
A. Vullo (*School of Computer Science and Informatics University College Dublin*)
G. Pollastri (*School of Computer Science and Informatics University College Dublin*)

Short Abstract: Residue contact maps are an important intermediate step towards de novo prediction of protein structure. The problem of accurately predicting contact maps from primary sequences is still unsolved. We focus on the problem of learning to predict physically realisable residue contact maps by directly incorporating rules encoding protein structural principles.

Long Abstract:

Protein topology representations such as residue contact maps are an important intermediate step towards de novo prediction of protein structure. Although improvements have occurred over the last years, the problem of accurately predicting residue contact maps from primary sequences is still largely unsolved. Among the reasons for this are the unbalanced nature of the problem (with far fewer examples of contacts than non-contacts) and the formidable challenge of capturing long-range interactions in the maps.

In order to mitigate the intrinsic difficulty of mapping one-dimensional input sequences into two-dimensional outputs, protein contact maps are inferred by modelling a set of independent tasks, one task being the prediction of the contact for a single pair of residues. Although the resulting problem turns out to be greatly simplified, this approach does not take into account the global nature of the problem, so that joining the outputs of all residue pairs often corresponds to a non-physically realisable map. This means that the structure derived from its contact map violates basic facts observed in real protein structures (e.g. the maximum number of contacts per amino acid) or in the worst case, that the residues could not be embedded into the three dimensional (3D) euclidean space.

We focus on the problem of learning to predict physically realisable residue contact maps. How to directly incorporate rules encoding protein structural principles into a learning framework remains an open issue. Here we present the results of a straightforward approach: we predict contact information by our state-of-the-art algorithm and then train a set of second stage multi-layered perceptrons as a filter mapping predicted patterns of contacts into actual outcomes. The mapping is implemented according to information related to physical realisability and observed in the predicted contact maps.

Predicted contacts are taken from our state-of-the-art contact map prediction system XXStout [1] that uses information from multiple alignment profiles, predicted secondary structure, solvent accessibility and contact density. XXStout's predictions contain errors such as the patterns of contacts between secondary structure elements are not very similar to those found in real contact maps, and that it assigns too many contacts to some AAs. Also contact order [2] distribution in predicted maps has a different shape to the one observed in real maps.

The filtering stage is implemented by training two different multi-layered perceptrons to filter different positions of the contact map. This is because the rules governing contact probability are likely to be different for positions near the main diagonal (mainly made by backbone atoms, reflecting secondary structure) than for positions away from it (i.e. contacts mainly occur between the side chains of amino acids placed in different helices or strands). For each pair of residues (i, j), the input features of both learners are derived by taking a square window of predicted contact probabilities centered on (i, j) and predicted secondary structure in three states, as given by our state-of-the-art system Porter [3]. To include information related with long-range contacts and map's physical realisability, these feature vectors are augmented with an encoding of the number of amino acids in contact with both residue i and j, the number of amino acids in contact with both residues and their respective contact order. Given these inputs, the position dependent multi-layered perceptrons learn to output the correct probability of contact for any given position taken into account.

We perform experiments to validate the ability of the system to recover correct contact information and to model maps with higher chance of physical realisability. To measure this, beside standard performance measures like precision, recall and their harmonic mean F1, we compare several distributions in real, predicted and filtered contact maps (Â-square test). The physical features considered are the number of contacts per amino acid, the number of contacts with both amino acids of each pair in contact and the number of contacts between residues in the same secondary structure element. The results indicate two facts: (1) filtered maps have slightly worse F1 values compared to predicted maps; (2) despite this, filtered maps feature a significant improvement in the distributions of physical related and long-range contact information.

Given these initial promising results, we plan to continue including more physicality related measures as those cited in [4], to finally develop a map quality measurement which will reflect not only correctly predicted contacts, but also map physicality.

References

[1] A. Vullo, IanWalsh, and G. Pollastri. A two-stage approach for improved prediction of residue contact maps. BMC Bioinformatics, 7:180, 2006.

[2] K. W. Plaxco, K. T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol., (277):985-994, 1998.

[3] G. Pollastri and A. McLysaght. Porter, a new, accurate server for protein secondary structure prediction. Bioinformatics, 21(8):1719-1720, 2005.

[4] Y. Shao and C. Bystroff. Predicting interresidue contacts using templates and pathways. Proteins, (53):497-502, 2003.