HIGH-THROUGHPUT COMPARATIVE MODELLING OF PROTEIN STRUCTURES BY MACHINE LEARNING

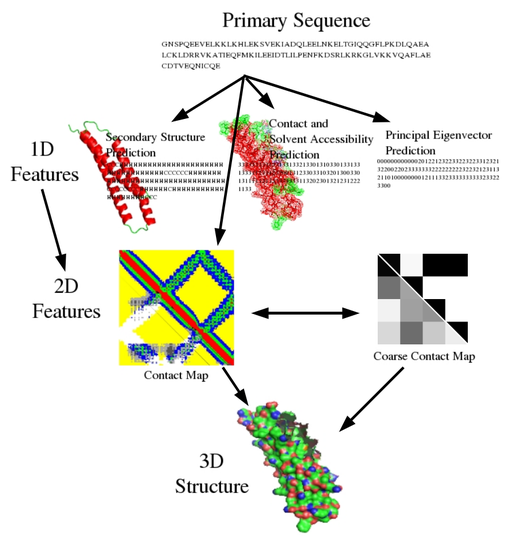
Clíona Roche, Davide Baú, Alberto J. Martin, Catherine Mooney, Alessandro Vullo, Ian Walsh, Gianluca Pollastri
{cliona.roche|davide.bau|albertoj|catherine.mooney|alessandro.vullo|ian.walsh|gianluca.pollastri}@ucd.ie

School of Computer Science and Informatics
University College Dublin
Belfield, Dublin 4, Ireland

Abstract

Of the over 4 million protein sequences currently known, only about 10% are human annotated, while for fewer than 2% has the three-dimensional structure been experimentally determined. Attempts to predict protein structure from primary sequence have been carried out for decades by an increasingly large number of research groups. Although the goal is far from being achieved in the field of *ab initio* prediction (where proteins have no resemblance to any structure in the PDB), *comparative modelling* (where templates do exist in the PDB) predictions are becoming more accurate. We propose to design, implement and test new algorithms for the prediction of protein structural features that incorporate template information from the Protein Data Bank (PDB) where available. The quality of these predictions will be benchmarked and made available in publicly available on-line large scale databases. It is hoped that these resources may become a fundamental source of information for scientists and may allow the development of new bioinformatics methods. Below we describe Distill - a suite of servers for the prediction of protein structural features.

Distill (http://distill.ucd.ie) is a suite of servers for the prediction of protein structural features: secondary structure; relative solvent accessibility; contact density; backbone structural motifs; residue contact maps at 6, 8 and 12 Angstrom; coarse protein topology[1−6]. The servers are based on large-scale ensembles of recursive neural networks and trained on large, up-to-date, non-redundant subsets of the Protein Data Bank. Together with structural feature predictions, Distill includes a server for prediction of Cα traces.

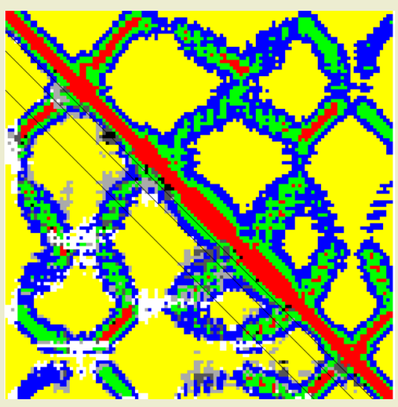


All structural feature predictors are based on single- or dual-layer Recursive Neural Network architectures for Directed Acyclic Graphs (DAG RNNs). One-dimensional feature predictors (i.e. those mapping the primary sequence into a sequence of the same length) are based on 1D DAG RNNs, while contact and distance map predictors are based on 2D DAG RNNs. Secondary structure, solvent accessibility and distance map predictors are provided structural information about PDB templates as a further input, when templates are available.

Multiclass Contact Maps

A multiclass contact map is an quantised version of the distance matrix (distances among the atoms in a protein), defined as follows:

- For a protein of *N* residues, the multiclass contact map is a *N x N* symmetric matrix **S**, whose elements S_{ij} are arrays of length *c*, where *c* = number of classes
- Two residues *i* and *j* are said to be in the class c_x whenever the mutual distance between their C_α atoms is within the distance boundaries defined for that class.



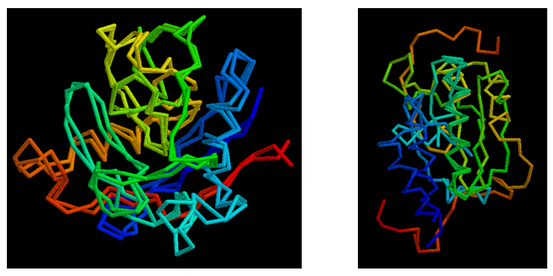


Fig. 1. Examples of reconstruction. Real structure superimposed to the predicted one for CASP targets (CM and FR_A-NF respectively) T0290 (left, 173 amino acids, TM score = 0.8846) and T0354 (right, 130 amino acids, TM score = 0.3084)

Acknowledgements

CR's work is supported through an IRCSET/Microsoft postgraduate scholarship 2006-2009. Distill's development was supported by Science Foundation Ireland grants 04/BR/CS0353 and 05/RFP/CMS0029, a UCD President's Award 2004, and an Embark Fellowship to AV from the Irish Research Council for Science, Engineering and Technology.

Reconstruction Algorithm

Goal: Quickly reconstruct draft protein structures for relatively short sequences (L $\leq$ 200)
Proteins are described at a coarse level as their backbone Cα trace

Bootstrap:

- Generation of an initial physically realisable configuration with a self-avoiding random walk and explicit modelling of predicted helices. A random structure is generated by adding Cα positions one after the other the whole backbone is represented

Search:[7]**:**

- Refinement of the initial bootstrapped structure by global optimisation of a pseudo-energy function using local moves and a simulated annealing protocol
- A randomly chosen point *i* is displaced (crankshaft move) **(a)**
- Secondary structure elements are displaced as a whole, without modifying their geometry **(b)**

CASP7 Results

Targets	*Ranked* 1^{st}	*Best submitted*
TM score		
CM easy	0.65840	0.66548
CM hard	0.40505	0.42811
FR_H	0.24443	0.25443
FR_A-NF*	0.25573	0.27730
GDT		
CM easy	0.50866	0.51550
CM hard	0.30451	0.32338
FR_H	0.19810	0.20754
FR_A-NF*	0.25632	0.27619

Table 1. TM score and GDT for CASP7 targets. *Ranked* 1^{st}: predicted structure ranked as first by our ranker. *Best submitted*: actual best submitted reconstruction (due to a glitch, we used the 2005 version of the PDB database for CASP predictions).

Targets	*Ranked* 1^{st}	*Best submitted*
TM score		
CM easy	0.68833	0.69497
CM hard	0.43766	0.45888
FR_H	0.25102	0.25543
FR_A-NF	0.25723	0.28248
GDT		
CM easy	0.53464	0.54169
CM hard	0.33330	0.35040
FR_H	0.20289	0.20546
FR_A-NF	0.25631	0.27903

Table 2. Same as Table 1, but using an updated version of the PDB database (last updated before the beginning of CASP experiments).

*Four of our FR_A-NF target predictions have been ranked in the top 10.

Current and Mid Term Research

Implementing a database of predicted protein structures:

- Designing the architecture of the database, the interface with the current system and the web interface.
- Benchmarking the current system on a small scale in order to :
  - Develope a reliability index to rate each prediction.
  - Establish which version of the pipeline is most stable.
- Benchmarking the current system on a large scale.
- Publishing the database publicly on the internet.

Future Research

- Analysing the current pipeline to see which areas require improvement.
- Implementing a number of comparative modeling algorithms to add to the existing pipeline.

References

1. G. Pollastri, A. McLysaght. "Porter: a new, accurate server for protein secondary structure prediction" Bioinformatics, 21(8):1719-20, 2005
2. A. Vullo, I. Walsh, G. Pollastri. "A two-stage approach for improved prediction of residue contact maps" BMC Bioinformatics, 7:180, 2006
3. D. Baú, A. J. M. Martin, C. Mooney, A. Vullo, I. Walsh, G. Pollastri. "Distill: A suite of web servers for the prediction of one- ,two- and three-dimensional structural features of proteins" BMC Bioinformatics, 7:402, 2006
4. C. Mooney, A. Vullo, G. Pollastri. "Protein Structural Motif Prediction in Multidimensional $\phi - \psi$ Space leads to improved Secondary Structure Prediction" Journal of Computational Biology, in press
5. A. Vullo, O. Bortolami, G. Pollastri, S. Tosatto. "Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines" Nucleic Acids Research, 34:W164-W168
6. G. Pollastri, A. Vullo, P. Frasconi, P. Baldi. "Modular DAG-RNN Architectures for Assembling Coarse Protein Structures" Journal of Computational Biology, 13:3, 631-650, 2006
7. M Vendruscolo, E Kussell, and E Domany. Recovery of protein structure from contact maps. Folding and Design, 2:295-306, 1997
8. J. Platt (2000) Probabilistic outputs of support vector machines and comparison to regularised likelihood methods. MIT press, Cambridge, MA
9. S. Vucetic, Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L.M. Iakoucheva, M.S. Cortese, J.D. Lawson, C.J. Brown, J.G. Sikes et al. (2005) Disprot: a database of protein disorder, Bioinformatics, 21, 137-140