Neural Network Pairwise Interaction Fields for Protein Model Quality Assessment

Alberto J.M. Martin, Alessandro Vullo, and Gianluca Pollastri

School of Computer Science and Informatics and
Complex and Adaptive Systems Laboratory
University College Dublin, Belfield, Dublin, Ireland
{albertoj,alessandro.vullo,gianluca.pollastri}@ucd.ie

Abstract. We present a new knowledge-based Model Quality Assessment Program (MQAP) at the residue level which evaluates single protein structure models. 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. We also attempt to extract local quality from global quality. The model allows fast evaluation of multiple different structure models for a single sequence. In our tests on a large set of structures, our model outperforms most other methods based on different and more complex protein structure representations in both local and global quality prediction. The method is available upon request from the authors. Method-specific rankers may also built by the authors upon request.

1 Introduction

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]. Moreover most protein structure prediction methods produce many reconstructions for any one protein and being able to sift out good ones from bad ones is often the key to the success of a method.

Several different potential or (pseudo-)energy function have been developed with the aim of mapping a three-dimensional structure into its “goodness” or “native-likeness”. These can be roughly divided into physics-based and knowledge-based, with some amount of overlap. The former are based on physico-chemical calculations [2, 3, 4, 5, 6], while the latter are based on statistics obtained from a training set of known structures (also termed “decoys”) typically generated by computational methods. These structures or decoys can be represented in different ways, embedded or not into a 3D lattice, in order to reduce the conformational space, and potentials can rely on more or less detailed representations of a structure. As simplicity in the representation increases (e.g. if each residue is represented as a single point or sphere), evaluation speed increases since fewer interactions have to be taken into account. By contrast the reliability of the evaluation tends to decrease when details are stripped from a structure [7, 8, 9, 10].
However, there may be other advantages in the use of simplified representations, such as reduced sensitivity to small perturbations in conformations and the ability to take into account complex effects that cannot be described separately [11], plus a reduced sensitivity to inaccuracies and uncertainties in the data (crystal contacts, high R-factors, etc.).

Knowledge-based potentials are not only used to rank models, but for other aims such as to drive 3D predictions and for model refinement [5, 6, 12, 13, 14]; fold recognition [15]; to place side chain and backbone atoms [16, 17]; as mentioned, to select native structures from sets of decoys [7, 8]; to predict stability of proteins [18, 19, 20]; or even to predict residue residue contact maps [21, 22, 23].

Several different representations of protein structures have been used in knowledge-based potentials. They can be classified in: 1) single point/sphere representations of each residue; 2) two or more points for each residue; 3) full atom models. In the first group each residue is usually represented by its $C_\alpha$ atom [24, 25, 26, 27, 28, 29] or by their $C_\beta$ atom [7, 8, 15, 18, 29, 30, 31], pseudo-$C_\beta$, side chain (SD) centre of mass or SD centroids [32]. In group 2) each residue is represented by two or more points, but fewer than its number of atoms; these points range from only two (e.g. $C_\alpha + C_\beta$, $C_\alpha + \text{pseudo-}C_\beta$ [33]), backbone atoms and some sort of representation of side chains like the $C_\beta$ [19], different rigid-body blocks/fragments [10]. In group 3) are many different potentials [7, 19, 20, 30, 31, 32, 36, 37]. These depend on different reference states and on different physico-chemical assumptions, whereas others are learnt from examples, typically via Machine Learning algorithms [28, 35, 38, 39].

There is a further group of potential functions - those based on clustering of many different structure predictions, and consensus methods. These usually outperform single model evaluation methods [1, 32, 40, 41]. Consensus methods make use of several different potentials to know the quality of the prediction [14, 38, 39]. The potentials are joined in a (possibly weighted) average. Clustering methods use similarities among high numbers of models obtained either from different methods (in this case they are hard to apply outside the CASP experiment environment (http://predictioncenter.org/), as many of the methods and their predictions are not easily available outside CASP), or to rank high numbers of reconstructions made by a single method [35, 42, 43].

The method presented here only uses information obtained from the $C_\alpha$ trace and the sequence of residues associated to it. The main advantage is that the quality of a model is assessed based on its overall topology, rather than based on local details. Moreover, there is no need to model backbone and side chain atoms before evaluating a structure, which allows many more $C_\alpha$ traces with different conformations to be produced. From $C_\alpha$ traces it is possible to model backbone and side chain atoms fairly accurately [16, 17], but this may be more computationally expensive than predicting several conformations of the protein structure as simple $C_\alpha$. If $C_\alpha$ traces can be evaluated effectively, backbone and side chains may be modelled only for those that are deemed to be accurate.