Adding energy calculations may help machine learning better predict protein structures

Wenfa Ng

Department of Chemical and Biomolecular Engineering, National University of Singapore,

Email: ngwenfa771@hotmail.com

Abstract

Machine learning has been utilized for predicting many types of variables in image processing, clinical diagnosis as well as board games through supervised learning based on a training dataset. Theoretically, the data driven approach of machine learning can be applied to many areas of scientific research and is only limited by the quality of the input data. One area where machine learning has increasingly been applied is the field of structural biology. Specifically, various machine learning methods including deep learning has been used in predicting the structure of proteins based on amino acid sequence. If the approach is validated for a variety of different classes of proteins, the method could aid the resolution of the protein folding problem, which hitherto could not be solved by energy calculations and molecular dynamics simulation. At heart in structural prediction of proteins is the determination of the correlation between amino acid sequence and protein folding pattern. Essentially, there are hidden rules and patterns in the highly complex dataset linking protein amino acid sequence and structure, which potentially could be uncovered by machine learning algorithms. Possibility thus exist in using the solved protein structures available in Protein Data Bank as well as the amino acid sequence of each protein for input into various machine learning algorithms to uncover features useful for protein structural prediction. Specifically, this approach sought to utilize the ability of machine learning tools to glean hidden relationships between protein amino acid sequence and specific structural motifs in a protein structure. Through this mapping, a correlation chart could be formulated where particular sequence motif is linked to specific protein fold. Although mutations could be present in a sequence motif previously linked to a structural fold, machine learning tools are generally tolerant of such imperfections and could still link a specific structural fold to an amino acid sequence motif. This holds two important implications: (i) machine learning-based protein structural prediction is inherently faster than *ab initio* molecular simulations, and (ii) structures predicted by machine learning may be coarse and unable to help answer questions on functional consequence of specific motifs. The latter is particularly important because the ultimately aim of structural biology is to arrive at structures able to help functional studies through linking structure to function. Perhaps, machine learning algorithms could be complemented by energy calculations derived by modelling small structural folds in three-dimensional space. Adding energy calculations would thus provide an additional layer of constraints to the final machine learning model, which may manifest as refinements to the structural picture emanating from artificial intelligence-based protein structure prediction. Overall, the application potential of machine learning in protein structure prediction has been explored, but many protein structures that manifest from such algorithms remain coarse and in need of structural refinement. Nevertheless, the fast speed and data driven approach of machine learning tools offers a simpler and more accessible path to computer-based prediction of protein structure compared to traditional molecular simulation. Adding

energy calculations to the training dataset may be one avenue in which machine learning-based protein structure prediction could be refined and help lend better protein models that informs functional studies in enzymology and beyond.

Keywords: machine learning, protein structural prediction, energy calculations, amino acid sequence, structural fold,

Subject areas: biochemistry, structural biology, bioinformatics, computational biology, biophysics,

Conflicts of interest

The author declares no conflicts of interest.

Funding

No funding was used in this work.