Qeios

Peer Review

Review of: "Evaluation of Molecular Docking by Deep Learning and Random Forests: A Hybrid Approach Based on Pseudo-Convolutions"

Congzhou Sha[1]

1. Medical Scientist Training Program, Pennsylvania State University, United States

The authors present a method for predicting if two proteins interact from their (genome) sequences. I have several comments that may be helpful in improving the quality of this manuscript.

First, the introduction is well-written, drawing attention to the monopoly that affluent countries have on the pharmaceutical industry and research, and that improvements in automation and machine learning may enable high-quality research to be done in the setting of limited resources.

On the technical side, my first comment is that there is no quantitative comparison to existing methods (e.g., ClusPro, AlphaFold 3). These existing tools are easily accessible to the public, and I think it would be valuable to at least reproduce the benchmarking of those tools on the specific datasets that the authors used, if such benchmarking exists.

My next comment is that there is not adequate justification for the pipeline shown in Figure 3, which preprocesses the input data. The approach is very similar to existing methods. The "pseudo-convolution" shown is essentially what has been incorporated as the attention mechanism of AlphaFold for multimeric structure prediction and has been explored in other deep learning articles (e.g., doi: 10.1186/s12859-021-04111-w). If the authors' intended message is that deep learning is unnecessary and that the simpler random forests are sufficient to perform this task, these arguments should be specifically made, and supporting data comparing to existing methods should be included.

Is there, for example, a mathematical justification for the pseudo-convolution encoding and why a simpler encoding (just the two sequences concatenated together) would not suffice? My suspicion is that just performing step #1 of Figure 3 would result in similar performance using random forests, as long as the individual trees are made sufficiently deep and/or the forest is sufficiently wide. Possible mathematical justification would be that the “pseudo-convolution” makes the

 doi.org/10.32388/X99801

“interactions” between the two sequences self-evident; however, the authors destroy this spatial structure by flattening the co-occurrence matrix in step #4 of Figure 3 for computing the pseudo-convolution.

As the authors note, a significant limitation of the datasets included is the class imbalances. For highly biased datasets, the receiver operating characteristic is not suited for analysis, and at the very least, the precision-recall curve should be plotted, with a comparison to classification via random chance. For example, the sensitivity and specificity are largely unchanged across all the random forests (Table 6), and 205 of 5153 (<4%) structures had interactions in the training set (Table 4), which represents severe class imbalance. It is also best practice to include the ROCs and PR curves in addition to quoting the AUCs. The authors do include appropriate cross-validation; however, there is no exploration of regularization. The fact that the sensitivities and specificities do not differ over a large range of forest sizes (from 100 trees to 1000 trees) indicates to me that the method is highly overparametrized/overfit. The ROCs and PRs should be shown for both training and validation sets to get an idea of how much overfitting has occurred.

On the biophysical side, all proteins will “interact” to some extent. The significance of these interactions can be determined by estimating the free energy of the bound and unbound states through molecular dynamics simulations and through experimental studies of the dissociation constant. Which method was used as the ground truth for “interacting” vs “non-interacting”? Additionally, specific proteins can be predicted to interact or not based on a basic understanding of biophysics. Proteins that are intrinsically disordered are unlikely to bind stably to other proteins, whereas highly structured proteins with classic beta sheets and alpha helices are well-known to have this interaction potential. The training/validation sets were not stratified by any such properties, and it is unclear if there is an advantage of the pseudo-convolution/random forest method over simpler models with fewer parameters (e.g., a multilinear regression on the percentage of hydrophobic residues/prolines/glycines/etc.).

There is not enough technical detail in the methods or code to exactly reproduce the results of this paper. Please include a code and data availability statement with a link to the code if possible.

Declarations

Potential competing interests: No potential competing interests to declare.