Kalpa Publications in Computing

Volume 14, 2023, Pages 21–23

Proceedings of V XoveTIC Conference. XoveTIC 2022

Capabilities of Approaches for Ab Initio Protein Structure Prediction

Juan Luis Filgueiras and José Santos

CITIC (Centre for Information and Communications Technology Research), Department of Computer Science and Information Technologies, University of A Coruña (Spain)
juan.filgueiras.rilo@udc.es, jose.santos@udc.es

Abstract

This paper presents a discussion of alternatives for protein structure prediction, showing the advantages and problems of recent alternatives based on deep learning and approaches based on energy optimization of protein solutions. A SARS-CoV-2 protein is included to exemplify the results.

1 Introduction

The structure of proteins largely determines their function, hence the great importance of determining their three-dimensional native structure. For this, traditional laboratory methods, such as X-ray crystallography and nuclear magnetic resonance, are expensive and time-consuming. Therefore, computational Protein Structure Prediction (PSP) methods attempt to close the gap between the number of proteins with known sequence (order of millions) and the number of solved proteins with known structure (about 180,000 in the PDB database [5]).

In the most difficult and challenging alternative of PSP, called ab initio, only the primary sequence information of the protein (its amino acid sequence) is used. This is based on Anfinsen's dogma [1], which states that the native structure is determined only with the primary sequence information and corresponds to the minimum Gibbs free energy. Consequently, an alternative in ab initio PSP is the use of search methods that try to discover the structure with minimum energy, once a protein representation and energy models are established. The problem is that PSP energy landscapes are high-dimensional and full of local minima. Thus, evolutionary computing search or optimization methods were intensively used, given their global search in multidimensional and multimodal energy landscapes.

In this line, our previous research used memetic algorithms (MAs) with protein atomic models, using the protein representation and energy model of Rosetta [7] (one of the most widely used software environments in PSP and protein design). Our *HybridDE* MA [9] combines the global search of Differential Evolution [6] with the local search provided by the protein fragment replacement technique, where the latter can locally refine protein structures maintained in the genetic population. Furthermore, given the inaccuracies of the Rosetta energy model, which provides a deceptive energy landscape in which the energy minimum need not correspond to the native structure, the crowding niching method was integrated into the MA (*CrowdingDE*)

[9][10]. This incorporation allows obtaining energy-optimized structures but, at the same time, with structural diversity, with the aim of discovering structures close to the native conformation.

As a different alternative, in recent years, the use of deep learning architectures for PSP has provided successful results in many proteins. For example, *RoseTTAFold* [2] and DeepMind's *AlphaFold* [8] provided predictions with higher accuracy on many proteins with respect to energy minimization approaches. The high capability of the deep learning methods is based on two aspects: i) the input information is given by the Multiple Sequence Alignment (MSA) of the target protein, which provides homologous protein information useful to know where there are possible amino acid contacts in the three-dimensional structure; ii) the training of the prediction models with a large number of known structures, the complete set of structures solved in the PDB [5] with the aforementioned PSP systems.

However, these deep learning-based methods do not work properly for proteins with poor MSA information. Moreover, their predictions may present flaws, such as amino acid conflicts, which require refinement based on energy minimization approaches. Therefore, here, building on the initial work published in [3], we present an example to show the pros and cons of both PSP alternatives (energy minimization and deep learning), with a SARS-CoV-2 protein.

2 Results and discussion

Figure 1 includes an example with a protein of SARS-CoV-2 virus. Protein *orf8* has 104 amino acids. This protein has no homologous proteins in the PDB database [5]. Its MSA coverage is poor, even searching for more homologous sequences in different genetic databases (Fig. 1, left part). Consequently, the prediction confidence of *AlphaFold2* models is low, as seen in Figure 1, with the PAE (Predicted Alignment Error) graph (Fig. 1, center). The Predicted Alignment Error at position (x, y) corresponds to the expected position error at residue x, when the predicted and real structures are aligned on residue y. This PAE level of confidence is poor at many positions (areas with red color, corresponding to high expected error). *AlphaFold2* model with the best confidence was selected, since *AlphaFold2* has a stochastic component (dropout during inference) and it can be run to provide several models. Likewise, *RoseTTAFold* solutions have similar confidence measures (not shown).

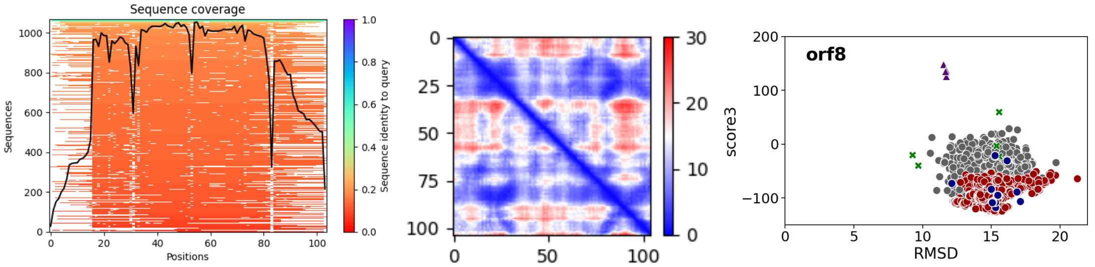

Figure 1: Left: MSA sequence coverage for SARS-CoV-2 protein *orf8*. MSA is input to *AlphaFold2* and *RoseTTAFold*. Center: PAE (Predicted Aligned Error) of the highest rated *AlphaFold2* model. Right: Energy (*score3*) vs. RMSD with different PSP approaches. Gray: Rosetta ab initio (run several times). Blue: *HybridDE* and Red: *CrowdingDE* (final optimized solutions). Green: *AlphaFold2* and Pink: *RoseTTAFold* (the 5 highest rated solutions).

Moreover, neither approach presents accurate solutions, as shown by the distribution of solutions in the energy vs. RMSD plot in Figure 1 (Fig. 1, right part). The x-axis corresponds to the RMSD (Root Mean Squared Deviation) distance between the superimposed structures (predicted conformations and the native structure). The distances are higher than 9 Å in all solutions, indicating that these are quite far from the native structure. The y-axis corresponds to the Rosetta energy (called *score3*) of the predicted solutions. The graph shows that the energy minimization-based approaches provide solutions with better energy compared to the deep learning-based approaches. This indicates the inaccuracies of the energy model, as the best solutions in energy terms do not correspond to those closest to the native structure. Finally, the evolutionary algorithm-based approaches perform better optimization (in energy terms) compared to the Rosetta ab initio protocol (based on a local search with fragment replacements), where, in this protein, *CrowdingDE* presents the solutions with the lowest energy and also a larger distribution of the optimized solutions with respect to *HybridDE*.

More examples of comparison of the different alternatives are included in [3], where, for most proteins, the deep learning approaches present more accurate solutions in terms of distance to the native structure. Moreover, visualizations of predictions with SARS-CoV-2 proteins can be seen in [4]. Future work is aimed at integrating the capabilities of both approaches in structure refinement, which is the main research line of the first author's PhD thesis.

3 Acknowledgments

This study was funded by Xunta de Galicia and the European Union (European Regional Development Fund - Galicia 2014-2020), with grants CITIC (ED431G 2019/01), GPC ED431B 2022/33 and IN845D-02 (funded by the "Agencia Gallega de Innovación", co-financed by Feder funds, supported by the "Consellería de Economía, Empleo e Industria" of Xunta de Galicia), and by the Spanish Ministry of Science and Innovation (project PID2020-116201GB-I00).

References

[1] C.B. Anfinsen. Principles that govern the folding of proteins. *Science*, 181(96):223–230, 1973.

[2] M. Baek, F. DiMaio, I. Anishchenko, and et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[3] J.L. Filgueiras, D. Varela, and J. Santos. Energy minimization vs. deep learning approaches for protein structure prediction. In *Proc. IWINAC 2022, Lecture Notes in Computer Science*, volume 13259, pages 109–118. Springer, 2022.

[4] Prediction results of the SARS-CoV-2 unsolved proteins. `https://www.dc.fi.udc.es/ir/in845d-02/SARS-CoV-2_protein_prediction/index.html`.

[5] Protein Data Bank. `http://www.wwpdb.org`.

[6] K.V. Price, R.M. Storn, and J.A. Lampinen. Differential evolution. A practical approach to global optimization, 2005.

[7] Rosetta system. `http://www.rosettacommons.org`.

[8] A.W. Senior, R. Evans, J. Jumper, and et al. Improved protein structure prediction using potentials from deep-learning. *Nature*, 577:706–710, 2020.

[9] D. Varela and J. Santos. Protein structure prediction in an atomic model with differential evolution integrated with the crowding niching method. *Natural Computing*, pages 1–15, 2020.

[10] D. Varela and J. Santos. Niching methods integrated with a differential evolution memetic algorithm for protein structure prediction. *Swarm and Evolutionary Computation*, 71:101062, 2022.