Protein structure model with high accuracy changes the game

Jihun Jeung[1,*]

[1]Gwangju Institute of Science and Technology (GIST College), 123 Cheomdan-gwagiro, Buk-gu, Gwangju 61005, Republic of Korea
[*]jeongjihun@gm.gist.ac.kr

ABSTRACT

Structrual bioinformatics is one of successful application in high performance computing (HPC). Protein structure is very complicated challenge and high performance computing elucidate this area with protein structure prediction model with high accuracy. In this paper, the basic concepts for structrual bioinformatics are discussed to understand the topic. Next, the overall procedure for template-free modeling and focus on evolutionary coupling and model refinement. At last, we will review the two cutting edge modeling, *AlphaFold2*[1] and *RoseTTAFold*[2] in view point of prediction accuracy and resource consumption.

Introduction

A protein is composed of amino acid sequences and has three-dimensional arrangement. Protein structure decides its function and roles in organism. The prediction of protein folding from amino acid sequence has been an important open question for more than 50 years.[3] Protein structure prediction and design is the results from multidisciplinary approach. It requires the understanding of energetics, dynamics, biology, and high-performance computing. As a result of continuous research, *AlphaFold2* and *RoseTTAFold* reaches the highest prediction accuracy in 14th Critical Assessment of protein Structure Prediction (CASP14). In this review, we will deal with basic concept of protein structure prediction and template-free protein structure prediction method. Then, we will compare *AlphaFold2* and *RoseTTAFold*.

Energy landscape and optimization algorithm

The structural prediction in structural bioinformatics is based on the hypothesis that a protein's folding conformation is completely encoded in its amino acid sequence.[4] Each conformation follows the energy landscape. The native state of a protein, its properly folded form, is the conformation with the lowest free energy.

The energy landscape in protein folding has multiple local minima. In other word, the challenge in structural bioinformatics is the wide space of potential conformations. In current technology, it is impossible to consider all potential conformations. Multiple levels of resolution are adopted in protein molecular modeling. At first, large-scale conformational sampling is conducted with high speed, efficient, and inaccurate coarse-grained energy function. Lowering the level of protein representation from all-atom to coarse-grained provide efficient implementation. Next, prior to all-atom modeling, all-atom resolution is recovered by backbone reconstruction, side chin reconstruction, and local optimization.[5] At last, all-atom modeling captures the modeling of amino acid side chains with a more time-consuming, more accurate, more sensitive to structural detail, and high-resolution atomistic energy function.

To find the folding conformation with lowest energy, three efficient sampling methods are used. Gradient-based optimization is the method in which the gradient of energy function with respect to the flexible degrees of freedom is calculated to find the gradient that follows the fastest way to local minimum. It is most popular optimization algorithm in machine learning. Although it is most efficient algorithm for local minimum, it does not guarantee to find the global minimum in non-convex loss function.[6]

Monte Carlo methods randomly select conformations with lower energy and sometimes escapes the local minima to find the global minima with low probability. Folding is simulated very fast with Monte Carlo dynamics.[7]

Molecular dynamic simulation is based on the Newton's laws of motion and potential energy function in protein folding such as kinetic energy, temperature, configuration energy, pressure, and so on. Molecular dynamics is the solution of the classical equations of motion for proteins in order to obtain the time evolution. Molecular dynamics program is (1) initialize the parameters, (2) compute the new forces, (3) solve the equation of motion, (4) sample, (5) test and increase time, (5) Repeat the computation as necessary and wrap up the simulation.

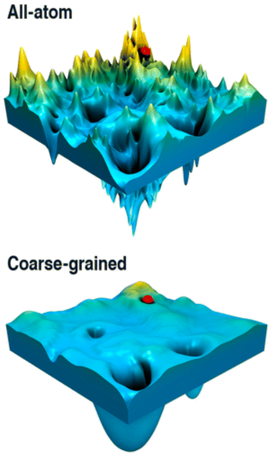

Figure 1. All-atom versus coarse-grained energy landscape. The figure illustrates the effect of the smoothening of the energy landscape in a coarse-grained model as compared to an all-atom model. The flattening enables efficient exploration of the energy landscape in search for the global minima, while avoiding traps in the local minima.

Protein Structure prediction method

There are two ways to predict protein structure. One is template-based modeling. Template-based modeling is to find the known protein structure that is similar with a protein of interest. It is knowledge-based prediction and requires the prior knowledge of protein structure. Template-based modeling is limited with roughly two-thirds of known protein families.

The other is template-free modeling. Template-free modeling is to predict protein structure without global structural similarity to a known protein. At first, a multiple-sequence alignment is constructed to predict local structure (e.g. local backbone structure and secondary structure) and to predict the spatial contact on the basis of covariation. Next, the 3D models are built by using previously predicted features with coarse-grained energy function and its optimization algorithm, distance geometry, and fragment assembly. At last, the 3D models are refined with all-atom energy function to better determine near-native prediction. (Figure 1)

Evolutionary coupling

Evolutionary pressure makes protein sequence keep favorable physical interactions among amino acid residues. The correlated mutations in protein sequences give this evolutionary information and it is used to estimate which pairs of residues are in contact. These pairs are referred as ‘evolutionary coupling’ and are predictive of functional sites.

To find evolutionary coupling, (1) build multiple sequence alignment for target sequence between many members of an evolutionarily related protein family. (2) Calculate covariance matrix (or co-occurrence frequencies) for all pairs of all amino acids. (3) Infer residue-residue co-evolution from statistical model and predict the physical contacts. (4) Get residue-residue distance constraints from predicted contacts. (Figure 3)[8,9]

Model refinement

Model refinement is the procedure that takes all-atom energy function, and it reaches the near-native conformation. Model refinement uses molecular dynamics simulation. Molecular dynamic simulation is the solution of the classical equations of motion in order to get insights on atomic or molecular level. The thermodynamic properties used in molecular dynamics are kinetic energy (Eqn. 1), temperature (Eqn. 2), configuration energy (Eqn. 3), pressure (Eqn. 4), speicfic heat (Eqn. 5), and so on.[10]

$$< K.E. > = < \frac{1}{2}\sum_{i}^{N} m_i v_i^2 > \tag{1}$$

$$T = \frac{2}{3Nk_B} < K.E. > \tag{2}$$

$$U_c = < \sum_{i} \sum_{j>i}^{N} \mathbf{V}(r_{ij}) > \tag{3}$$

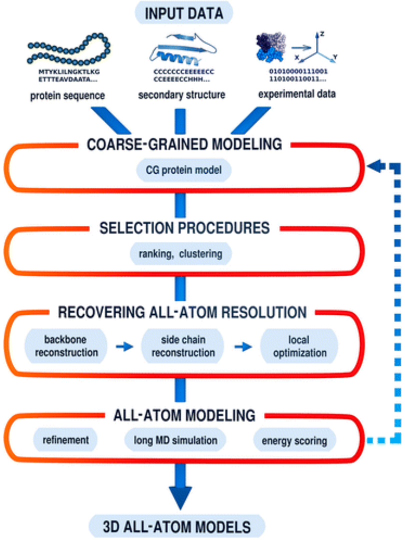

Figure 2. Typical multiscale modeling scheme that merges coarse-grained and all-atom modeling. In specific tasks, the resulting all-atom structures could be used as an input for the next stage of coarse-grained simulations. Other multiscale schemes are briefly discussed in the text.

$$PV = Nk_BT - \frac{1}{3} < \sum_{i=1}^{N-1} \sum_{j>i}^{N} \overrightarrow{\mathbf{r}_{ij}} \cdot \overrightarrow{\mathbf{f}_{ij}} > \tag{4}$$

$$< \delta(U_c)^2 >_{NVE} = \frac{3}{2} Nk_B^2 T^2 (1 - \frac{3Nk_B}{2C_v}) \tag{5}$$

The structure of molecular dynamics program is following :

- Initialization
- Compute the new forces
- Solve the equation for motion
- Sample
- Test and increase time
- End of the simulation

Emergence of protein structure prediction models with high accuracy

AlphaFold2 achieve high performance in 14th Critical Assessment of protein Structure Prediction (CASP14).[1,7] *AlphaFold2* is completely different model from *AlphaFold*[11]. Contrast to *AlphaFold*, *AlphaFold2* are the combination of computational molecular modeling and biological inference.

The model consists of two blocks of the network, Evoformer and Structure Module. Evoformer is the representation of protein structure prediction as graph inference problem in 3D space. Evoformer interprets the biological question as a graph. 'Structure Module' represents the geometry of backbone in amino acid (N-Calpha-C atoms). Compared the previous version, Alphafold, *AlphaFold2* includes the direct starting from multiple sequence alignments, an attention mechanism, and a two-track network architecture. An attention mechanism and two-track network architecture captures 1D sequence level and 2D distance map better than the one-track network architecture and convolution neural network.

The shortage of *AlphaFold2* is the slow speed. *AlphaFold2* spends 4.8 minutes for 256 residues, 9.2 minutes for 384 residues, and 18 hours for 2,500 residues. Also, *AlphaFold2* takes a huge memory space of GPU. A single GPU of V100 is

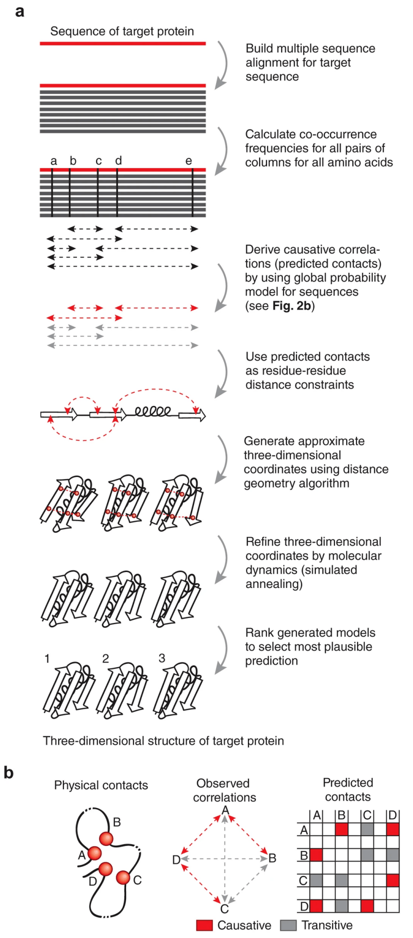
a
Sequence of target protein
Build multiple sequence alignment for target sequence
Calculate co-occurrence frequencies for all pairs of columns for all amino acids
a b c d e
Derive causative correlations (predicted contacts) by using global probability model for sequences (see **Fig. 2b**)
Use predicted contacts as residue-residue distance constraints
Generate approximate three-dimensional coordinates using distance geometry algorithm
Refine three-dimensional coordinates by molecular dynamics (simulated annealing)
Rank generated models to select most plausible prediction
1
2
3
Three-dimensional structure of target protein
b
Physical contacts
Observed correlations
Predicted contacts
A
B
C
D
Causative
Transitive

Figure 3. Workflow as implemented on the publicly available web server EVfold.org. The amino acid sequence of the target protein is used to perform a database search for putative structural homologs, with attention to the optimal cutoff in sequence similarity so that sufficient sequences are available yet they are not too far diverged to lose subfamily specificity. Minimally, hundreds of sequences are needed to derive plausible causative evolutionary couplings.

used to compute a 2,500-residue protein. Consider that the destination of protein structure prediction is to design a therapeutic protein and that most of antibody (one of common type of therapeutic protein) is more than 3,000 residues. Taking a lot of time and resource may be a barrier to design a new drug with *AlphaFold2*. In drug development, a lot of drug candidates are tested to find the most effective protein conformation. AlphFold2 has room for further improvement.

Interestingly, *AlphaFold2* was not publicly available until July 2021 and science community who got an inspiration from *AlphaFold2* designed and published another structure prediction model, *RoseTTAFold*.[2] The architecture of *RoseTTAFold* is two-track model augmented with a third parallel structure track for 3D backbone coordinates. Owing to the limitation of memory space, it was impossible to implemente 3D backbone coordinates as third track.

Even though *RoseTTAFold* does not show the good accuracy as much as that of *AlphaFold2*, *RoseTTAFold* shows higher average TM-score than all of protein prediction model except *AlphaFold2*. Moreover, *RoseTTAFold* is more efficient than *AlphaFold2* in respect to time and resource consumption. It requires 8GB of memory for over 400 residues protein. A several hours is enough to predict a protein with less than 400 residues.

Discussion

Opensource software accelerates the advance of HPC

The current achievement in structural bioinformatics such as *AlphaFold2* is the result of multidisciplinary collaboration and a legacy of open software. *AlphaFold2* merges the molecular dynamics from classical mechanics, the evolutionary inference from biology, and deep learning algorithm from computer vision. This merge in structural bioinformatics is impossible without opensource software including Rosetta[12, 13], pyRosetta[14], EVfold[9], and so on. A single research group can not cover every domain. The collaboration of multiple research group is necessary for outstanding result. One key idea in an opensource software leads to another idea to solve the challenges in structural bioinformatics. Finally, structural bioinformatics community reaches to the protein structure prediction model with high accuracy. AlphaFold2 does not reach its performance in a day. It owe to the accumulated opensource softwares and their cleaver idea to solve the challenges.

Predicting phenotype from omics data is next game for HPC

One of the biggest question in biology is here: Can we predict a phenotype from omics data? After the emergence of next generation sequencing (NGS), we have a variety kind of omics data on the level of genomics, transcriptomics, proteomics, epigenomics, and metabolomics. Even though a lot of omics data have been accumulated, predicting phenotype with these data is not easy. Genome-scale modeling is one of an attractive modeling method to predict phenotype. Genome-scale model use a reconstructed map and calculate a flux of reaction.[15]. Genome scale model give a window to predict a few phenotype change in prokaryotic organism.[16] Combination of genome-scale model and deep learning[17] may also become another 'game changer'. The current limitation of genome scale model is the deficiency of reliable data for enzyme reactivity. To apply deep learning, genome scale model need a reliable proteomics data such as K_{cat} and K_m. One cleaver idea is to predict these constant accurately using deep learning method and apply the predicted value for genome scale modeling[18, 19]. Until now, deep learning model that uses structural information of each enzyme is not implemented. A deep learning model that uses a protein structure prediction model with high accuracy like *AlphaFold2* could be fascinating future research topic.

References

1. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* DOI: 10.1038/s41586-021-03819-2 (2021).

2. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* DOI: 10.1126/science.abj8754 (2021). https://science.sciencemag.org/content/early/2021/07/19/science.abj8754.full.pdf.

3. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697, DOI: 10.1038/s41580-019-0163-x (2019).

4. Anson, M. L. & Mirsky, A. E. PROTEIN COAGULATION AND ITS REVERSAL : THE PREPARATION OF INSOLUBLE GLOBIN, SOLUBLE GLOBIN AND HEME . *J. Gen. Physiol.* **13**, 469–476, DOI: 10.1085/jgp.13.4.469 (1930). https://rupress.org/jgp/article-pdf/13/4/469/1186584/469.pdf.

5. Kmiecik, S. *et al.* Coarse-grained protein models and their applications. *Chem. Rev.* **116**, 7898–7936, DOI: 10.1021/acs.chemrev.6b00163 (2016). PMID: 27333362, https://doi.org/10.1021/acs.chemrev.6b00163.

6. Géron, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems* (O'Reilly, Beijing Boston Farnham Sebastopol Tokyo, 2019), second edition edn.

7. Kolinski, A. & Skolnick, J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* **18**, 338–352, DOI: 10.1002/prot.340180405 (1994).

8. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**, 1072–1080, DOI: 10.1038/nbt.2419 (2012).

9. Marks, D. S. *et al.* Protein 3d structure computed from evolutionary sequence variation. *PLOS ONE* **6**, 1–20, DOI: 10.1371/journal.pone.0028766 (2011).

10. Hollingsworth, S. A. & Dror, R. O. Molecular dynamics simulation for all. *Neuron* **99**, 1129–1143, DOI: https://doi.org/10.1016/j.neuron.2018.08.011 (2018).

11. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710, DOI: 10.1038/s41586-019-1923-7 (2020).

12. Leaver-Fay, A. *et al.* Rosetta3. In *Methods in Enzymology*, vol. 487, 545–574, DOI: 10.1016/B978-0-12-381270-4.00019-6 (Elsevier, 2011).

13. Leman, J. K. *et al.* Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* **17**, 665–680, DOI: 10.1038/s41592-020-0848-2 (2020).

14. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691, DOI: 10.1093/bioinformatics/btq007 (2010). https://academic.oup.com/bioinformatics/article-pdf/26/5/689/561368/btq007.pdf.

15. Fang, X., Lloyd, C. J. & Palsson, B. O. Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nat. Rev. Microbiol.* DOI: 10.1038/s41579-020-00440-4 (2020).

16. Du, B., Yang, L., Lloyd, C. J., Fang, X. & Palsson, B. O. Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in escherichia coli. *PLOS Comput. Biol.* **15**, 1–21, DOI: 10.1371/journal.pcbi.1007525 (2019).

17. Zampieri, G., Vijayakumar, S., Yaneske, E. & Angione, C. Machine and deep learning meet genome-scale metabolic modeling. *PLOS Comput. Biol.* **15**, 1–24, DOI: 10.1371/journal.pcbi.1007084 (2019).

18. Heckmann, D. *et al.* Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nat. Commun.* **9**, 5252, DOI: 10.1038/s41467-018-07652-6 (2018).

19. Heckmann, D. *et al.* Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. *Proc. Natl. Acad. Sci.* **117**, 23182–23190, DOI: 10.1073/pnas.2001562117 (2020).