

TSUBAME 共同利用 令和 5 年度 学術利用 成果報告書

深層学習を用いた分子動力学シミュレーションの高速化
Acceleration of molecular dynamics simulation using deep learning林 周斗
Shuto Hayashi東京医科歯科大学 難治疾患研究所
Medical Research Institute, Tokyo Medical and Dental University
<https://www.shimamlab.info/>

タンパク質が持つ機能を解析するため、分子動力学シミュレーションによりタンパク質立体構造のダイナミクスを解析する研究が広く行われている。しかしながら、分子動力学シミュレーションには大きな計算コストを必要とするという問題点があり、長時間・大規模なシミュレーションを行うにはスーパーコンピュータを用いても膨大な時間を要する。そこで本課題では、分子動力学シミュレーションを高速かつ高精度に模倣する深層ニューラルネットワークを開発した。前年度に開発したモデルの汎化性能を上げるため、16 生物種から得られた計 36,555 種類のタンパク質に対して分子動力学シミュレーションを実行し、学習データセットを構築した。

To analyze the functions of proteins, molecular dynamics simulations are widely used to study the dynamics of protein structures. However, molecular dynamics simulations have the drawback of requiring significant computational resources, and even with the use of supercomputers, conducting long and large-scale simulations can be time-consuming. In this project, we developed a deep neural network that can mimic molecular dynamics simulations with high speed and accuracy. To enhance the generalization performance of our model developed in the previous year, molecular dynamics simulations were performed on a total of 36,555 proteins from 16 species to build a training dataset.

Keywords: molecular dynamics simulation, deep learning, E(3)-equivariant neural network, graph neural network, protein structure

背景と目的

タンパク質が持つ機能はその立体構造によって決定づけられており、タンパク質の立体構造を精査することはその機能を知る上で重要な手がかりとなる。これまで X 線結晶構造解析、核磁気共鳴分光法、電子顕微鏡などの実験手法により、さまざまなタンパク質の立体構造情報が解析されてきた。しかしながら、これらの手法によって得られる立体構造は一時刻点における静止像であるため、タンパク質立体構造の動的変化を捉えることはできない。実際、薬学分野においては、薬・標的タンパク質間の相互作用を見極めるために立体構造の動的変化を解析することが重要視されている。

これまでタンパク質立体構造の動的変化を解析する研究が数多くなされてきたが、それらの中でも特に広く用いられているのが分子動力学シミュレーションである。しかしながら、分子動力学シミュレーションには大きな計算コストを必要とするという問題点があり、スーパーコンピュータを用いたとしても現実的な時間内で計算で

きるのは長くともマイクロ秒～ミリ秒スケールのシミュレーションまでである。

そこで本課題では深層学習をベースとしたデータ駆動型のアプローチを用いることで分子動力学シミュレーションを高速に行う手法を開発することを目的とする。

概要

本課題ではタンパク質の全原子座標、速度、力、原子種を入力として、一定時間後の全原子座標、速度、力を高速・高精度に予測する深層ニューラルネットワーク(LAMDA)を学習する。学習データセットとして、実際のタンパク質分子動力学シミュレーションデータを用いる。ネットワークの構築、および学習には PyTorch を用いる。

結果および考察

本年度は、前年度に開発したモデルの汎化性能を上げることを目指した。まず、AlphaFold Protein

Structure Database^[1,2]に登録された 16 生物種、計 326,172 種類のタンパク質配列に対して、MMseqs2^[3]を用いてクラスタリングを行った(表 1)。得られた 126,649 クラスターのうち、クラスターサイズが 1 のものは以降の解析から除外した。結果として残った 36,555 クラスターのそれぞれから代表タンパク質を 1 つずつ選び、GROMACS^[4]を用いた分子動力学シミュレーションを実行した。

表 1: MMseqs2 を用いたタンパク質フィルタリング

生物種	クラスター リング前	クラスター リング後	フィルタ リング後
Arabidopsis	27,434	7,722	3,050
Nematode worm	19,694	11,852	1,927
C. albicans	5,974	3,452	817
Zebrafish	24,664	8,054	3,524
Dictyostelium	12,622	8,743	1,120
Fruit fly	13,458	8,303	1,313
E. coli	4,363	3,299	543
Soybean	55,799	17,393	5,761
Human	23,391	4,155	2,726
M. jannaschii	1,773	1,491	195
Mouse	21,615	3,257	2,751
Asian rice	43,649	23,739	3,945
Rat	21,270	4,123	2,905
Budding yeast	6,039	3,259	831
Fission yeast	5,128	3,035	577
Maize	39,299	14,772	4,570
計	326,172	126,649	36,555

まとめ、今後の課題

LAMDA の汎化性能を向上させるため、計 36,555 種類のタンパク質に対して分子動力学シミュレーションを実行し、学習データセットを構築した。今後はこのデータを用いて LAMDA の学習を行い、速度と精度の検証を行う。

参考文献

- [1] Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-589.
- [2] Varadi M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50(D1):D439-D444.
- [3] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026-1028.
- [4] Abraham MJ, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1:19-25.