

令和 6 年 5 月 3 日現在

機関番号：32660

研究種目：学術変革領域研究(B)

研究期間：2021～2023

課題番号：21H05157

研究課題名（和文）ベイズ統計学を組み入れた次世代構造モデリングプラットフォームの構築と応用

研究課題名（英文）Development of the next-generation structure modeling platform incorporating Bayesian inference

研究代表者

森 貴治 (Mori, Takaharu)

東京理科大学・理学部第一部化学科・准教授

研究者番号：90402445

交付決定額（研究期間全体）：（直接経費） 25,500,000円

研究成果の概要（和文）：生命現象を解明するためには、タンパク質複合体の構造とダイナミクスを調べる必要がある。しかしながら、一般的に実験データにはノイズやエラーが多く含まれ、特に巨大なタンパク質複合体の場合、実験データからの正確な立体構造モデリングが難しくなる。本研究課題では、ベイズ推定を用いて実験データを分子動力学（MD）計算に組み込む MELD 法を MD 計算プログラム GENESIS に導入し、構造モデリングプラットフォームを構築した。これにより、複数の実験データを拘束条件として用いても、ノイズやエラーを自動的に除外しながらタンパク質複合体の立体構造を効率的に予測することが可能になった。

研究成果の学術的意義や社会的意義

タンパク質の立体構造を予測することは、生物物理学における挑戦的課題の一つである。特に近年では、クライオ電顕やAFM、NMR、クロスリンク質量分析データなど、複数の実験手法から得られる情報を統合して巨大タンパク質複合体の構造を予測する、“Integrative structural biology”が注目されている。本研究課題において、こうした問題に対する新たなソフトウェア基盤の構築を行った。本成果が、実験科学者にとって利用しやすいソフトウェア環境を提供するだけでなく、それを通して医学や薬学において重要なタンパク質の立体構造予測にも貢献できると期待している。

研究成果の概要（英文）：To elucidate biological phenomena, it is necessary to examine the structure and dynamics of protein complexes. However, experimental data typically contain a lot of noise and errors, making accurate 3D structural modeling challenging, especially for large protein complexes. In this research project, we introduced the MELD method, which incorporates experimental data into molecular dynamics (MD) calculations using Bayesian inference, into the MD simulation program GENESIS to construct a structural modeling platform. This allows for the efficient prediction of protein complex structures while automatically excluding noise and errors when using multiple experimental data as constraints.

研究分野：理論生物物理学

キーワード：分子動力学 タンパク質複合体 ベイズ推定

1. 研究開始当初の背景

細胞内の生命現象を理解するためには、各現象の過程で形成されるタンパク質複合体の立体構造やダイナミクスを調べるのが重要であり、近年、このような解析を目的とした実験技術が著しく進歩した。例えば、クライオ電子顕微鏡を用いることでタンパク質の立体構造を高解像度で決定できるようになり、また、高速原子間力顕微鏡（高速 AFM）を用いることでタンパク質の動きをリアルタイムで観察できるようになってきた。また、クロスリンク質量分析実験からは、タンパク質中での近接するアミノ酸配列の距離情報がわかり、このような実験データは、タンパク質複合体における構成因子の配向や位置の予測にも利用される。これらの構造解析実験や生化学実験を組み合わせることで、従来困難であった巨大タンパク質複合体の立体構造を正確に決定できるようになってきた。しかしながら、一般的に実験から得られる観測データには、タンパク質分子の揺らぎや構造多型、ノイズ、エラーなどの様々な情報が含まれる。メガダルトン級の巨大タンパク質複合体を対象として構造やダイナミクスを解析しようとした場合、系の自由度が高く、複雑な相互作用が存在するため、タンパク質単体に対する解析よりも実験データの解釈が困難となる。このような問題を克服するためには、実験データから効率的にタンパク質の立体構造を予測するための新たな方法論が必要となる。

2. 研究の目的

実験データからタンパク質の立体構造をモデリングする手法として、例えば、クライオ電顕フレキシブルフィッティングなどのように、実験データを拘束条件とした MD 計算を行う方法が広く用いられる。これはシミュレーションで用いるエネルギー関数に、理論値と実験値とのズレをバイアスとして加える方法である。しかしながら、実験データには、タンパク質の動的性質に起因する確率的な構造情報や、ノイズやエラーなどによる不確かな情報が混在するため、すべての実験データを拘束条件として用いると、実験データへのオーバーフィッティングによって実際の構造から歪んだ予測構造が得られる可能性がある。従って、エラーやノイズを含む実験データをいかにして合理的に拘束条件として取り込むかが重要である。本研究では、メガダルトン級タンパク質複合体の立体構造モデリングを見据え、実験データを合理的に拘束条件として取り込み、立体構造を精密に予測するための新しい方法論およびソフトウェアの開発を目指した。具体的には、分子動力学計算プログラム GENESIS に、クライオ電顕や高速 AFM、クロスリンク質量分析データなどのさまざまな実験データを「ベイズ推定」を介して MD 計算に取り込む、データ駆動型シミュレーション法を導入し、次世代構造モデリングプラットフォームを構築することを目指した。

3. 研究の方法

本研究では、ベイズ推定に基づいて実験データを拘束条件として取り入れる方法の 1 つである MELD 法を用いた。ベイズ推定に基づく、分子構造 \mathbf{x} が与えられたときに、実験データが D であるときの確率は、条件付き確率 $P(D|\mathbf{x})$ で求められ、MELD 法では系のエネルギー関数を以下のように定義して実験データに関する拘束を取り入れる。

$$\begin{aligned} E_{Total} &= -k_B T \ln p(\mathbf{x}|D) \\ &= E_{FF} - k_B T \ln p(D|\mathbf{x}) \\ p(D|\mathbf{x}) &\sim \prod_i^N \exp[-\beta E_i^{Restraint}(\mathbf{x})] \end{aligned}$$

ここで N はいわゆるカットオフで、「全拘束条件のうち N 個の拘束条件を利用する」ことを意味しており、各拘束条件をそのエネルギーに従って並び替え、小さいものから順に N 個選ぶ。

$$E_1^{Restraint} \leq E_2^{Restraint} \dots \leq E_N^{Restraint} \dots$$

このような拘束条件を用いることにより、実験データに一致する拘束条件は系に付加され、逆に実験データから大きく逸脱した拘束条件、すなわち、エラーあるいはノイズとみなされるものは自動的に棄却されることになる。本研究では、実験データとしてクライオ電顕およびクロスリンク質量分析データを扱えるように GENESIS に MELD 法を導入した。

GENESIS control file

```
[INPUT]
groupfile = ../3_AIC02/lwv_conv_eg.top
grocefile = ../3_AIC02/lwv_conv_eg.gro
meldfile = ../4_MaskData/DataSet1_200_80

[OUTPUT]
mdfild = md_ded # DVD trajectory file
pdbfile = md.pdb # PDB file

[ENERGY]
forcefield = RESIDCO # (CHARMM,AMBER,KAGO,KRHO,AMBER,GROMBER,GROMARTIN)
electrostatic = CUTOFF # [CUTOFF,PMX]
electrostatic = CUTOFF # Debye-Huckel model
rg_parallelizet_svr = 15.0 # Neighborlist distance

[DYNAMICS]
integrator = VVDR_CG # velocity-Verlet propagation
nsteps = 2000 # number of MD steps
timestep = 0.01 # timestep size (ps)
emout_period = 50 # energy output interval
troutout_period = 50 # trajectory output interval
logupdate_period = 10 # pairlist update interval
iseed = 1

[CONSTRAINTS]
rigid_bond = NO # don't apply constraints

[ENSEMBLE]
ensemble = NVT # Canonical ensemble
epsilon = LAMMPS # Langevin thermostat
temperature = 300 # simulation temperature
gamma_t = 0.01 # thermostat friction parameter

[BOUNDARY]
type = NBOC # periodic boundary condition

[SELECTION]
group1 = anrCA # restraint group 1

[RESTRAINTS]
restraints = 1 # number of functions
function1 = DR # restraint function type
constants = 1200 # force constant
select_index1 = 1 # restrained groups

[EXPERIMENTS]
emfit = YES # EM fit
emfit_target = ../2_EMMap/Invt.ccp4 # Density map file (*.map *.mrc *.ccp4 *.sit)
emfit_resolution = 2.5 # resolution parameter of the simulated map
emfit_tolerance = 0.001 # Tolerance for error
```

MELD file

```
[DISTANCE]
threshold = 80
7 35 0 0 0.1 30
7 128 0 0 0.1 30
7 158 0 0 0.1 30
7 169 0 0 0.1 30
7 189 0 0 0.1 30
7 204 0 0 0.1 30
30 120 0 0 0.1 30
30 135 0 0 0.1 30
35 120 0 0 0.1 30
35 135 0 0 0.1 30
35 141 0 0 0.1 30
35 158 0 0 0.1 30
35 234 0 0 0.1 30
47 120 0 0 0.1 30
47 252 0 0 0.1 30
47 825 0 0 0.1 30
120 128 0 0 0.1 30
120 135 0 0 0.1 30
120 141 0 0 0.1 30
```

図1 クライオ電顕フレキシブルフィッティングとクロスリンク質量分析実験データに基づく距離拘束を MELD 法により組み合わせたタンパク質の立体構造予測を行うための GENESIS のインプットファイル例
(左)MD 計算の詳細な条件を定義したコントロールファイル、(右)距離拘束を定義した入力ファイル

4. 研究成果

図1にクライオ電顕像とクロスリンク質量分析データに基づき、MELD法を用いてタンパク質の立体構造を予測するための GENESIS のコントロールファイル例を示す。本導入における重要なポイントは、従来のクライオ電顕フレキシブルフィッティングあるいは通常の MD 計算用のコントロールファイルの仕様は変更せず、MELD 法で用いる拘束条件を付加する際は、[INPUT] セクションに MELD 用入力ファイルの情報を一行追加するだけである。入力ファイルを meldfile と定義し、これに距離や角度、二面角、位置に関する拘束条件を記入するようにした。さらにフラットボトムポテンシャルを導入し、力の定数や参照値をフレキシブルに定義できるようにした。meldfile 中の threshold は実験データのうちの何%を利用するかを意味しており、これは前頁のカットオフ N に相当する。MELD 計算は、MD 計算中に拘束条件の並び替えを毎ステップ行うため、クイックソートおよびハイブリッド MPI/OpenMP 並列計算を導入することで、多量の拘束条件があったとしても高速に計算できるように工夫した。

導入した MELD 法をテストするために、カルシウムイオンポンプを対象として、人工的に生成したクロスリンク質量分析データとクライオ電顕像を用いてフレキシブルフィッティングによる立体構造予測の精度を検証した。なお、クロスリンクデータには 20% のエラー (overlength ペア) を混入させており、クライオ電顕像の解像度は 5 Å とした。クロスリンクデータの threshold を 0%, 70%, 80%, 90% と変え、それぞれ 20 回の計算を実行したところ、電顕情報だけ (0%) では、低解像度のために正しく構造が予測できなかったケースがあるのに対して、クロスリンクデータを追加した場合、特に実際のエラー率に近い threshold = 80% を用いたケースでは、20 回の計算すべてにおいてほぼ正解構造に辿り着いた (図2)。一方、70% は 0% と違いがほとんどなく、90% は overfitting のためにほとんどの予測構造が歪んでいた。以上の結果から、クライオ電顕フィッティングに効率的にクロスリンクデータの拘束を取り込むことができ、実験データ中のエラー率がおおよそわかれば、MELD 法にてそのようなエラーを自動的に除外しながら構造を精密にモデリングすることが可能であることが示唆された。

本研究で開発した MELD 法は今後、GENESIS 最新版に導入して無償にて公開する予定である。また、チュートリアルを作成して公開し、実験研究者への普及も図っていく。

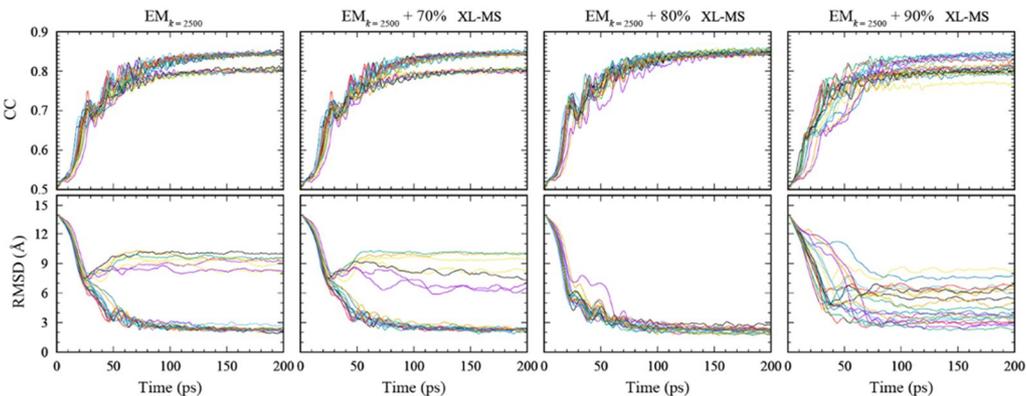


図2 クライオ電顕フレキシブルフィッティングに、80% のエラー情報を含むクロスリンク質量分析 (XL-MS) データを追加したときに得られる CC (理論マップと実験マップとの相関係数) と RMSD (正解構造からのずれ) の時間変化。

5. 主な発表論文等

〔雑誌論文〕 計5件（うち査読付論文 5件/うち国際共著 0件/うちオープンアクセス 3件）

1. 著者名 Dokainish Hisham M, Re Suyong, Mori Takaharu, Kobayashi Chigusa, Jung Jaewoon, Sugita Yuji	4. 巻 11
2. 論文標題 The inherent flexibility of receptor binding domains in SARS-CoV-2 spike protein	5. 発行年 2022年
3. 雑誌名 eLife	6. 最初と最後の頁 e75720
掲載論文のDOI（デジタルオブジェクト識別子） 10.7554/eLife.75720	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -
1. 著者名 Yagi Kiyoshi, Re Suyong, Mori Takaharu, Sugita Yuji	4. 巻 72
2. 論文標題 Weight average approaches for predicting dynamical properties of biomolecules	5. 発行年 2022年
3. 雑誌名 Current Opinion in Structural Biology	6. 最初と最後の頁 88 ~ 94
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.sbi.2021.08.008	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -
1. 著者名 Kohga Hidetaka, Mori Takaharu, Tanaka Yoshiki, Yoshikaie Kunihiro, Taniguchi Katsuhide, Fujimoto Kei, Fritz Lisa, Schneider Tanja, Tsukazaki Tomoya	4. 巻 30
2. 論文標題 Crystal structure of the lipid flippase MurJ in a “squeezed” form distinct from its inward- and outward-facing forms	5. 発行年 2022年
3. 雑誌名 Structure	6. 最初と最後の頁 1088 ~ 1097
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.str.2022.05.008	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -
1. 著者名 Yu Isseki, Mori Takaharu, Matsuoka Daisuke, Surblys Donatas, Sugita Yuji	4. 巻 45
2. 論文標題 SPANAs: Spatial decomposition analysis for cellular scale molecular dynamics simulations	5. 発行年 2023年
3. 雑誌名 Journal of Computational Chemistry	6. 最初と最後の頁 498 ~ 505
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/jcc.27260	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -

1. 著者名 Ikei Mai, Miyazaki Ryoji, Monden Keigo, Naito Yusuke, Takeuchi Azusa, Takahashi Yutaro S., Tanaka Yoshiki, Murata Keina, Mori Takaharu, Ichikawa Muneoyoshi, Tsukazaki Tomoya	4. 巻 22
2. 論文標題 YeeD is an essential partner for YeeE-mediated thiosulfate uptake in bacteria and regulates thiosulfate ion decomposition	5. 発行年 2024年
3. 雑誌名 PLOS Biology	6. 最初と最後の頁 e3002601
掲載論文のDOI (デジタルオブジェクト識別子) 10.1371/journal.pbio.3002601	査読の有無 有
オープンアクセス オープンアクセスとしている(また、その予定である)	国際共著 -

〔学会発表〕 計9件(うち招待講演 3件/うち国際学会 0件)

1. 発表者名 森貴治
2. 発表標題 Structure modeling of protein complex from experimental data using molecular dynamics simulation
3. 学会等名 第60回日本生物物理学会年会(招待講演)
4. 発表年 2022年

1. 発表者名 森貴治
2. 発表標題 分子動力学計算と実験データに基づくタンパク質複合体立体構造モデリング
3. 学会等名 第22回日本蛋白質科学会年会
4. 発表年 2022年

1. 発表者名 森貴治、寺師玄記、松岳大輔、木原大亮、杉田有治
2. 発表標題 Cryo-EM flexible fitting refinement with automatic error fixing for de novo protein structure modeling
3. 学会等名 第59回日本生物物理学会年会
4. 発表年 2021年

1. 発表者名 森貴治
2. 発表標題 MD 計算と AI 技術を組み合わせた実験データ解析
3. 学会等名 多次元相分離研究会 (招待講演)
4. 発表年 2023年

1. 発表者名 森貴治
2. 発表標題 シミュレーションによる生体分子のダイナミクス解析
3. 学会等名 第 5 回 超越分子システム 計算科学を使う勉強会 (招待講演)
4. 発表年 2024年

1. 発表者名 村岡俊佑、森貴治
2. 発表標題 大腸菌のタンパク質透過チャネルSecYの立体構造予測とMD計算による膜中での安定性の評価
3. 学会等名 第13回日本生物物理学会関東支部会
4. 発表年 2024年

1. 発表者名 佐藤克樹、森貴治
2. 発表標題 タンパク質透過装置Secトランスロコンの構造変化過程の解析
3. 学会等名 第13回日本生物物理学会関東支部会
4. 発表年 2024年

1. 発表者名 松原大都、森貴治
2. 発表標題 分子動力学計算に基づく巨大タンパク質複合体のクロスリンク質量分析データの解析
3. 学会等名 第3回「富岳」成果創出加速プログラム研究交流会
4. 発表年 2024年

1. 発表者名 森貴治
2. 発表標題 分子動力学計算による脂質フリッパーゼ MurJ の構造変化の解析
3. 学会等名 日本化学会第104春季年会(2024)
4. 発表年 2024年

〔図書〕 計0件

〔産業財産権〕

〔その他〕

-

6. 研究組織

氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
---------------------------	-----------------------	----

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------