令和５年度高性能汎用計算機高度利用事業

「富岳」成果創出加速プログラム
「生体分子シミュレータを基にした大規模推論システムの開発と応用」
成果報告書

令和６年５月３０日

国立大学法人　埼玉大学

松永　康佑

補助事業の名称

「富岳」成果創出加速プログラム
生体分子シミュレータを基にした大規模推論システムの開発と応用

体系的番号： JPMXP1020230119

１．補助事業の目的

生体分子シミュレータを活用して、実験データから逆問題を解き、分子構造やダイナミクスに関する大規模な統計的推論を実現する研究開発を行う。ここでは、「富岳」向けに最適化されている分子動力学法(Molecular Dynamics; MD)計算プログラム GENESIS に注目し、若手を中心とするメンバーで統計的推論に応用できるようにする。具体的には、GENESIS を推論システムにおける尤度計算部品とみなし、ベイズ推論・機械学習/AI と組み合わせて構造空間の探索範囲をうまく絞り込むことで、これまで不可能であった高次元サンプリングを実現し生命科学の新たな知見を得る。

２．令和５年度（報告年度）の実施内容

２－１．当該年度（令和５年度）の事業実施計画

(サブ課題 1) マルコフ状態モデルを介した構造ダイナミクスの推論（埼玉大・松永）

本課題は、大規模分子シミュレーションと高速原子間力顕微鏡(AFM)データを融合させるデータ同化を行い、天然変性タンパク質の構造遷移ダイナミクスの推論を行うことを目的とする。一般に MD シミュレーションで到達できる時間スケールと高速 AFM データの時間解像度にはギャップがあり、両者を同化することは困難であるが、これまでに松永らは両者の間に中間的なスケールのモデル(マルコフ状態モデル)を介することでデータ同化を行い構造ダイナミクスを推論することに成功してきた。しかし一方で、マルコフ状態モデルを構築するには MD シミュレーションで当該タンパク質がとり得る構造を網羅的にサンプリングする必要があり、適用できる分子系が限られてきた。令和 5 年度は、サロゲートモデルにより駆動された能動学習の開発を行う。特に、構造緩和が遅い構造ダイナミクスを能動的にサンプルする実装とテスト計算を行う。

(サブ課題 2) 遺伝子発現と共役するゲノム 3 次元構造動態の推論（京大・高田）

本課題は、大規模生体分子シミュレーションにより、ゲノム 3 次元構造に依存した遺伝子発現制御の分子機構の推論を行うことを目的とする。Hi-C 等のゲノムデータに基づいてヘテロなゲノム構造アンサンブルを推論するために、ベイズ推論ベースの metainference 法を発展させるとともに、マルチスケールシミュレーションを実施する。令和 5 年度は、Hi-C Metainference 法を用いた ES 細胞 Nanog 遺伝子座の複数の構造モデル構築を行う。次に得られたメゾスコピック構造モデルと ChiP-seq データを用いて残基分解能構造モデルを構築する。また、ES 細胞 Nanog 遺伝子座の代表構造モデルについての長時間シミュレーションとその解析を行う。

(サブ課題 3) クライオ電顕データからの構造アンサンブル推論（名大・Tama）
本課題は、クライオ電顕の画像データから生体分子の構造アンサンブルやサンプル不均一性についての情報を推論するための手法を開発し、実験データに応用することを目的とする。具体的には、クライオ電顕個々の画像を解析し、それぞれが捉えている生体分子構造を推論することによりサンプル内の構造の多様性に関する情報を得ることを目的とする。令和 5 年度は、まず既存の実験データによるアルゴリズムのパフォーマンスの検証を行う。プロトタイプとして実装している GENESIS コードの高度化とともに、「富岳」での動作確認と性能評価を行う。

(サブ課題 4)エラー・ノイズを含む実験データからの統合的立体構造推論（東理大・森）
本課題は、クライオ電顕実験データとクロスリンク質量分析実験データを統合的に組み合わせてタンパク質複合体の立体構造を予測するための技術基盤の構築と応用を行うことを目的とする。電顕像の解像度が低く、電顕像のみからは構成タンパク質の配向がはっきりしない場合、アミノ酸間距離情報を含むクロスリンク質量分析実験データがしばしば参照されるが、クロスリンク質量分析データは実験条件によっては 30%近くエラーが含まれることもあり手動モデリングによる主観が入る。この問題をベイズ推論をベースにした MELD 法の応用により解決することが目的である。令和 5 年度は、主に以下の３つについて取り組む。1) クライオ電顕とクロスリンク質量分析データを MELD 法、レプリカ交換法、全原子/粗視化モデルを組み合わせて解析できるように GENESIS を改良する、2)カルシウムイオンポンプをテスト系としてプログラムの検証を行う。3)ヌクレオソーム、転写因子の MD 計算を行い、実験データと比較する。

理研・小林らは、上記の課題と協力して、課題計算を「富岳」で有効にスケーリングさせるためのプログラムの高度化を行う。必要な場合には GENESIS への機能追加を行う。

明大・中村らは、上記の課題と協力して、推論を効率的に行うための機械学習/AI モデリングを行う。具体的には、課題 1 の松永らのサロゲートモデル構築を協力する。

(プロジェクトの総合推進)
プロジェクト全体の連携を密としつつ円滑に運営していくためのミーティングを、オンラインで定期的に開催する。また、年末にプロジェクト内の研究の進捗状況および成果の発表のためのワークショップ等を開催する。プロジェクトで得られた成果については学会発表等により、積極的に公表する。また、他のプロジェクトとの連携などにより、効率的・効果的な研究の推進を行う。

２－２．実施内容（成果）
(サブ課題 1) マルコフ状態モデルを介した構造ダイナミクスの推論（埼玉大・松永）
本課題は、大規模分子シミュレーションと高速 AFM データを融合させるデータ同化を行い、天然変性タンパク質の構造遷移ダイナミクスの推論を行うことを目的とする。一般に MD シミュレーションで到達できる時間スケールと高速 AFM データの時間解像度にはギャップがあり、両者を同化することは困難であるが、これまでに我々は両者の間に中間的なスケールのモデル(マルコフ状態モデル)を介することでデ

ータ同化を行い構造ダイナミクスを推論することに成功してきた。しかし一方でマルコフ状態モデルを構築するには、MD シミュレーションデータを射影して定義する状態がマルコフ性を満たすとは限らないこと、およびターゲットタンパク質がとり得る構造を網羅的にサンプリングする必要があることから、適用できる分子系が限られてきた。構造サンプリングや状態定義が難しい天然変性タンパク質のデータ同化を実現するために、令和 5 年度は、AI ベースのサロゲートモデルを介した次元縮約法を開発するとともに、構造緩和が遅い分子系に対して能動サンプリングによる網羅的構造サンプリングのテスト検証を行った。

MD シミュレーションデータからマルコフ状態モデルを構築するには、まずデータを次元縮約し、縮約された空間でクラスタリングして状態を定義して構築するが、定義した状態がマルコフ性を満たす保証はないという問題がある。この問題に対して、緩和モード解析や tICA が提案されてきたが、いずれも線形変換という限界があり、単純なドメイン運動として記述が困難である天然変性タンパク質への応用に関しては課題が残る。深層学習ベースの手法は非線形変換であるが、一般にはダイナミクスは考慮されていない。例えば Variational autoencoder (VAE)はデータを latent space 上でガウシアンに分布するように縮約するが、画像などの学習を前提としているため、静的な分布情報しか考慮していない。そこで、機械学習/AI 専門の明大・中村グループと連携して、ダイナミクスを考慮するように VAE を拡張した(逐次変分オートエンコーダの一種と捉えることができる) tsVAE ならびに tsTVAE 手法を開発した。tsVAE と tsTVAE は、latent space の prior として、分布の代わりに単純なマルコフ過程のひとつである Ornstein-Uhlenbeck (OU)過程が満たすべきモーメントを課すことで、複雑なタンパク質ダイナミクスを単純な確率過程へ縮約する学習を行うことができる。開発した手法を、alanine-dipeptide や chignolin の MD シミュレーションデータで検証し、latest スペースにおけるダイナミクスがよりマルコフ性を持つこと、ならびに disentangled な表現が得られており解釈性が高いことを示した（図 1-1、Ishizone et al. *J. Chem. Theory Comput.* **20,** 436-450 (2023))。

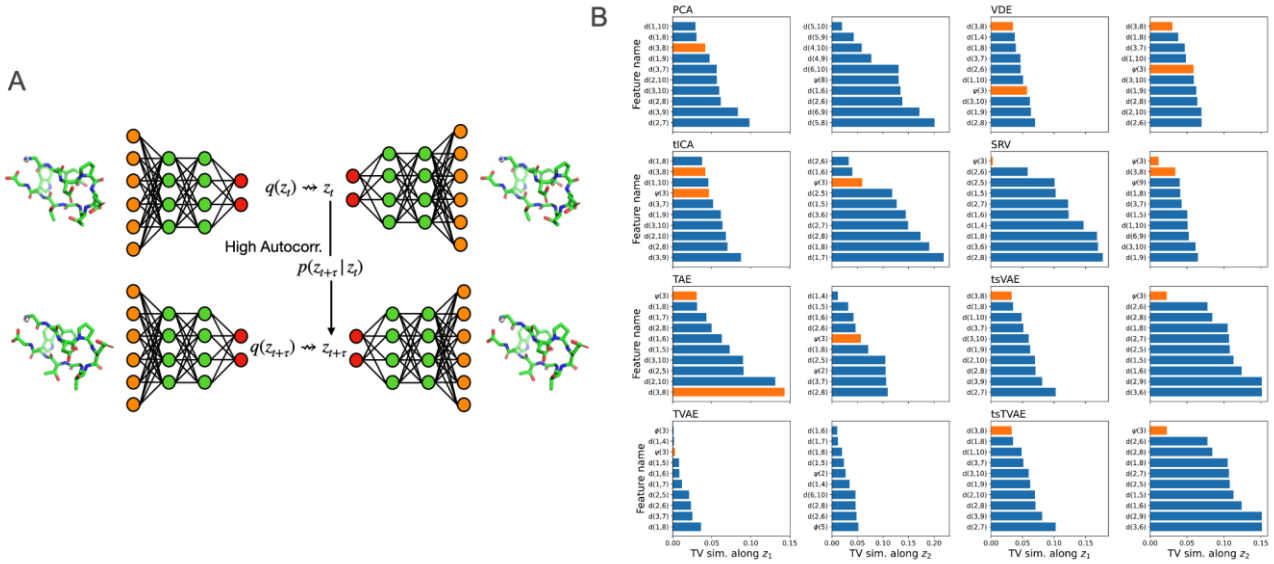

図 1-1: ts(T)VAE の概念と比較検証結果。(A) ts(T)VAE の概念図、異なる時間の座標を使って、latent space 上で遅い自己相関に相当する表現ができるように学習する。(B) Chignolin の MD シミュレーションで比較検証した結果。Disentanglement の評価指標である total variation similarity を様々な手法と比較している

(Ishizone et al. *J. Chem. Theory Comput.* **20,** 436-450 (2023)より引用)。

マルコフ状態モデルを構築するためにもう一つ重要な課題として、ターゲット分子の構造空間を事前に網羅的に構造サンプリングできていなければならない、という要件がある。小分子系やドメイン運動で記述できる低次元系では困難ではないが、構造空間が広大な天然変性タンパク質では構造空間の網羅的サンプリングは挑戦的な課題である。構造空間の効率的なサンプリングのために多くの enhanced sampling 手法が提案されているが、その中から本課題のターゲットに適しており、「富岳」上で GENESIS を用いて実装可能な手法である、能動学習法に基づいたサンプリングに取り組んだ。これは注目する低次元座標や縮尺された空間上で、獲得関数値等の評価を行い、その値に基づいて次の探索範囲の意思決定を行う。実装が簡単なことから Weighted ensemble や PACS-MD に倣って、複数のレプリカシミュレーションを行って、その結果をリサンプリングすることで次の探索範囲の絞り込みを行った。

能動学習に基づいたサンプリング法のターゲットとして、味覚受容体 t1r の全原子構造モデルを用いた構造サンプリングに取り組んだ。この系はヘテロダイマーであり、活性化構造(閉構造)が実験構造として得られているが、不活性状態がどのような構造をとっていて活性化構造へ至るのか、そのメカニズムが不明な系である。それまでの粗視化モデルシミュレーションや 1 分子計測実験から、中間状態として非対称構造をとることが示唆されていたが、エネルギー論の定量評価やその原因となる相互作用が不明であった。能動学習に用いる低次元座標としてダイマー間の距離(Cysteine-rich domain 間の距離)を用いて 100 レプリカシミュレーションのリサンプリングを「富岳」上で行うことで構造サンプリングを行った(前出の ts(T)VAE と同時期に開発を行っていたため、ここでは ts(T)VAE の latent space ではなくシンプルな距離座標を使った)。その結果 GPU を用いた brute-force では得られたなかった広大な構造空間をサンプリングすることができ、自由エネルギー地形を得ることができた (図 1-2)。その結果、1 分子計測実験と整合的な中間状態を再現することができ、更に片方の非対称構造のみしか遷移経路とならないことを明らかにし、非対称性を決定するインタフェース相互作用を調べることができた。現在同様の枠組みで、ts(T)VAE の latent space を使った Protein G 構造サンプリングを進行中である。

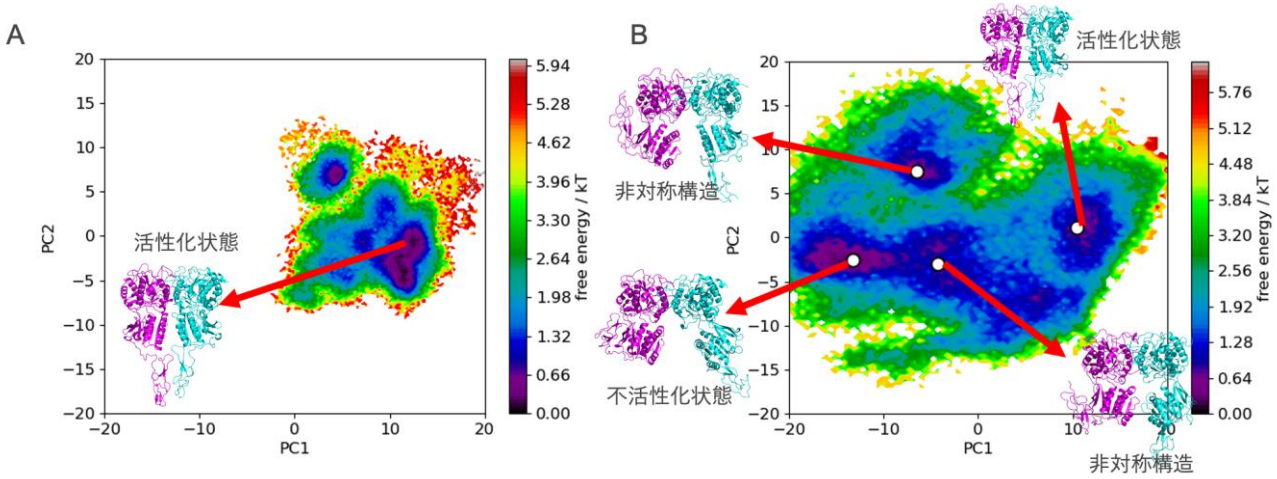

図 1-2: 味覚受容体の全原子 MD シミュレーションの能動学習の比較検証結果(自由エネルギー地形)。(A) GPU を用いた brute-force MD シミュレーションによる構造サンプリングの結果(2 μ 秒のシミュレーション時間のジョブを独立に 10 セット。GPU 10 枚を使用しておよそ五ヶ月間の計算) を。(B) 「富岳」上での能動学習による構造サンプリングの結果 (令和 5 年度前半の計算資源を利用)。

(サブ課題 2) 遺伝子発現と共役するゲノム 3 次元構造動態の推論（京大・高田）

本課題は、大規模生体分子シミュレーションにより、ゲノム 3 次元構造に依存した遺伝子発現制御の分子機構の推論を行うことを目的とする。Hi-C 等のゲノムデータに基づいてヘテロなゲノム構造アンサンブルを推論するために、ベイズ推論ベースの metainference 法を発展させるとともに、マルチスケールシミュレーションを実施する。

多細胞生物は多様な細胞種をもち、多様性は細胞種ごとに異なる遺伝子発現に起因する。これが高次生命現象の基礎となっている。細胞種依存的な転写制御において、エピゲノム制御およびそれと密接に関連したクロマチンの3次元折りたたみ構造の変化が主要な役割を果たす。本課題では哺乳類の胚性幹(ES)細胞に着目し、幹細胞とその分化で主要な機能を果たす遺伝子群の転写制御とクロマチン構造の関係を、細胞スケールのMDシミュレーションによって推論する。哺乳類ES細胞のコア遺伝子ネットワークは山中因子を含む3つの遺伝子、*Oct4, Sox2, Nanog*で構成され、ES細胞の多能性と自己複製能の維持に必須である。本課題ではマウス*Nanog*遺伝子座（図2-1）を本研究の標的とし、Micro-CおよびChIP-seqなどの実験データをもとに、その遺伝子座の折りたたみ構造と発現活性化機構を推論する。

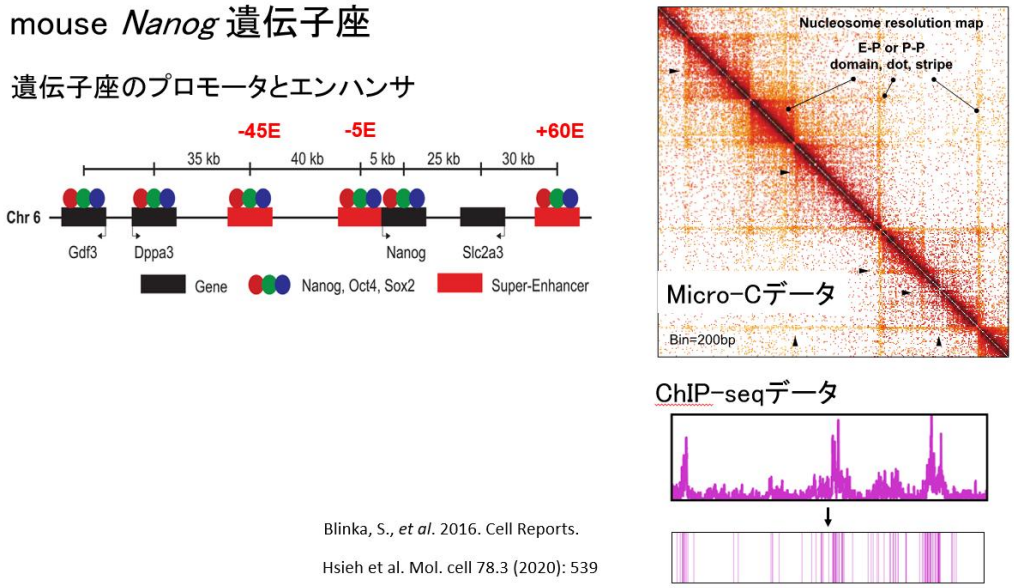

図2-1: 対象としたマウス *Nanog*遺伝子座の特徴と主な実験データ。（左上）*Nanog*遺伝子座200kb領域は、そのプロモータ、3つのスーパーエンハンサ（-45SE, -5SE, +60SE）、および関連する他の3つの遺伝子座（プロモータ）を含む。（右上）Micro－Cデータ。（右下）ChIP-seqデータの一例。

令和 5 年度は、Hi-C Metainference 法を用いた ES 細胞 *Nanog* 遺伝子座の複数の構造モデル構築を行った。Hi-C Metainference は、我々が開発した Hi-C（あるいは Micro-C）データに基づくベイズ推定法である。我々は昨年度までに、1kb 分解能ポリマーモデルを *Nanog* 遺伝子座に適用し、実験データを再現する構造アンサンブルを得ている（図 2-2A）。本年度、その構造アンサンブルから、マルコフ状態遷移モデルを構築し、大きく異なる 5 つのクラスタとその間の遷移速度に関する知見を得た（図 2-2B）。遺伝子座は多様でヘテロな構造の間を柔らかく動き回っている。さらに、分解能を上げてヌクレオソーム分解能のクロマチンモデルを適用し、より詳細な構造状態を特定することに初めて成功した（図 2-2C）。

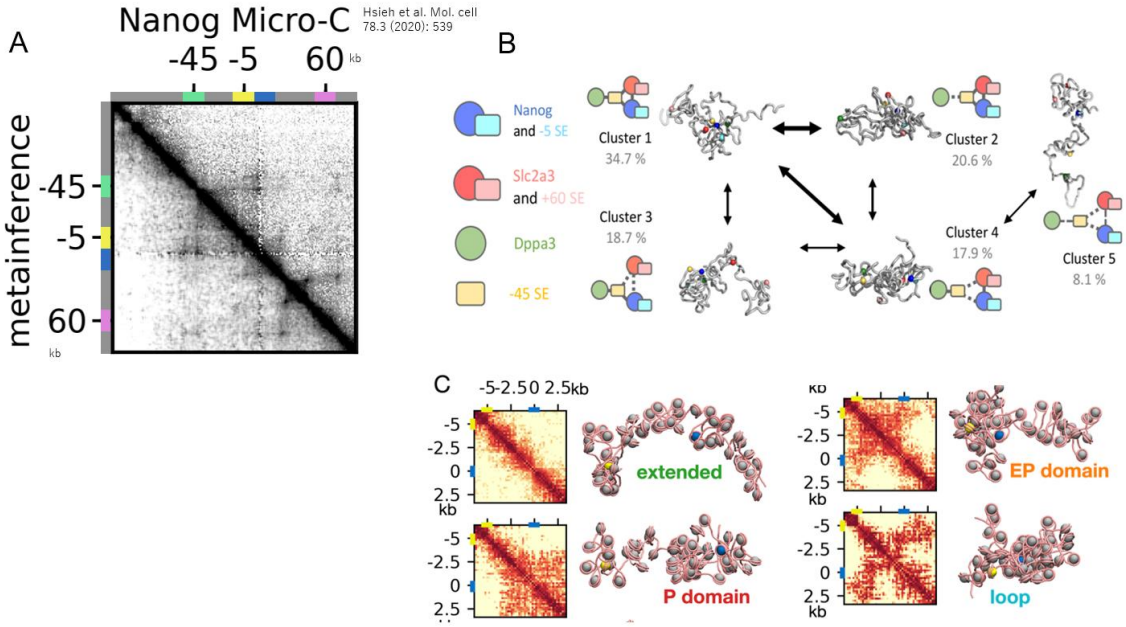

図2-2: Hi-C Metainference法を用いたマウス Nanog遺伝子座の複数の構造モデル。（A）Hi-C Metainference法により得られるコンタクトマップ（左下三角）は、実験データ（右上三角）を再現した。（B）構造アンサンブルから得られた5つの構造状態とその間の遷移速度。（C）ヌクレオソーム分解能モデルによるHi-C Metainference法の計算結果。

次に、得られたメゾスコピック構造モデルと ChiP-seq データを用いて残基分解能構造モデルを構築した。我々自身の先行研究により、メゾスコピックモデルから残基分解能構造モデルを構築するプロトコルを開発している。これを *Nanog* 遺伝子座の 5 つの代表的構造モデルに適用し、それぞれの残基分解能モデルを得た（図 2-3 左）。5 つのうち、コンパクトな構造ほど粒子同士の重なりや高分子の結び目を生じるリスクが高いことが分かった。これらを自動的に解消するための方法を工夫し、概ね自動的にリスクを回避した構造モデルを構築できる段階になった。構築した構造モデルには、DNA、コアヒストン、リンカーヒストンを含むクロマチン、RNA 合成酵素、メディエータ、コアクチベータ（p300, Brd4）、コア転写因子（Oct4, Sox2, Klf4, Nanog）が含まれる。

また、ES 細胞 *Nanog* 遺伝子座の代表構造モデルについての長時間シミュレーションとその解析を行った。令和 5 年度末までに 10^8 MD ステップの計算を完了し、初期の構造緩和、コアクチベータと転写因子の中規模な結合・解離を観察することが出来た。コンパクトな構造状態においてだけ、スーパーエンハンサと *Nanog* プロモータを繋ぐ分子ネットワークが見られ（図 2-3 右下）、遺伝子活性化機構の一端を観察しつつあると考えられる。

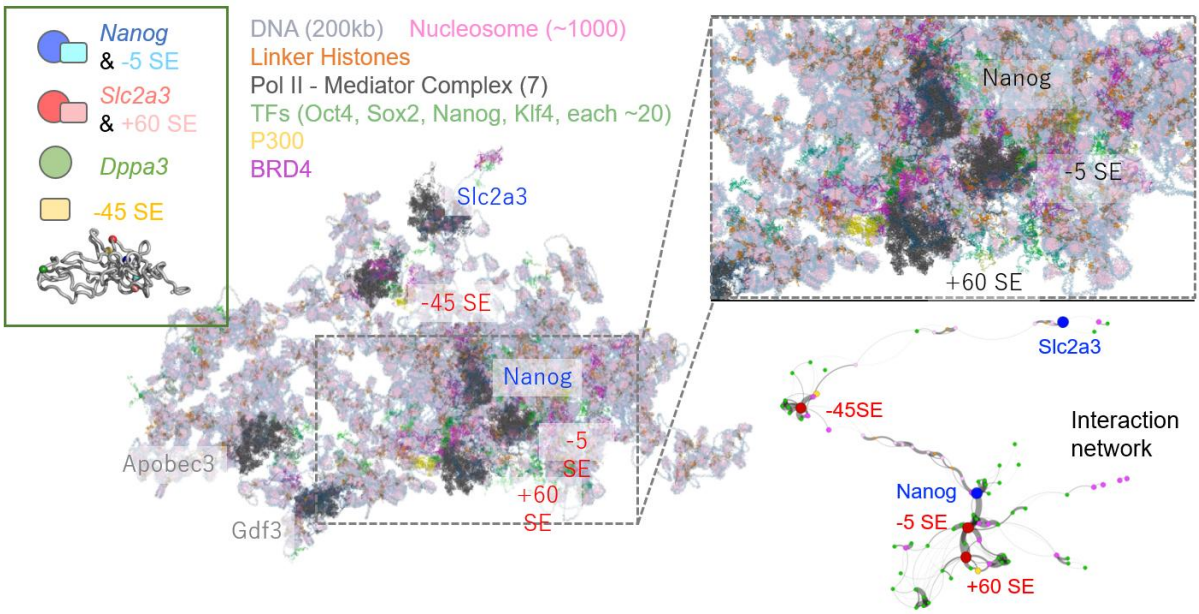

図 2-3: *Nanog* 遺伝子座の残基分解能モデルの一つ（構造 2）。DNA（灰色）、コアヒストン（ピンク）、リンカーヒストン（オレンジ）を含むクロマチン、RNA 合成酵素（黒色）、メディエータ（黒色）、コアクチベータ（p300（黄色）, Brd4（紫色））、コア転写因子（Oct4, Sox2, Klf4, Nanog、すべて緑色）。（右下）構造モデルにおける分子間の相互作用ネットワーク。3 つのスーパーエンハンサ(-45SE, -5SE, +60SE)とプロモータが連結している。

(サブ課題 3) クライオ電顕データからの構造アンサンブル推論（名大・Tama）

This project aims to investigate the continuous conformational variability and dynamics of biological macromolecules from large-scale cryo-EM single-particle 2D images using MD simulations employing the MDSPACE algorithm (Vuillemot et al., *J. Mol. Biol.* **435**, 167951 (2023)). The current research has two main goals. The first objective is to assess the accuracy of the fitting method. The second objective is to increase the efficiency of the software for conducting large-scale flexible fitting using the Fugaku supercomputer. This will be accomplished by modifying GENESIS and developing tools that automatically set up essential computational stages of MDSPACE on Fugaku.

3.1 Performance evaluation and parameter optimization of the MDSPACE method

We evaluated the fitting performance of different integrators and force fields in the MD simulation. As a test system for the evaluation purpose, we use a protein complex, glutamate dehydrogenase (GDH), an enzyme pivotal in glutamate amino acid metabolism, which facilitates the reversible transformation of glutamate into α-ketoglutarate (Figure 3-1). The GDH from *Thermococcus profundus*, in its unbound state, forms a homohexameric structure, with each unit comprising a nucleotide-binding domain (NBD) and a core domain. This enzyme demonstrates significant structural flexibility, characterized by the spontaneous movement of the NBD towards the core domain, thereby modulating its open and closed states.

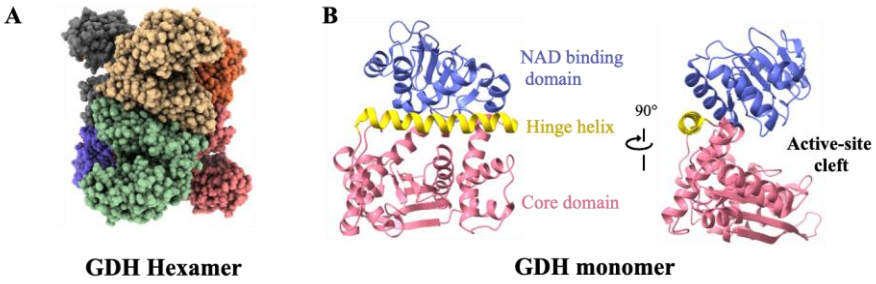

Figure 3-1: (A) Crystal structure of GDH hexamer (PDB: 1EUZ). The six monomers are colored differently. (B) One of the monomeric structures. The NAD-binding domain, hinge helix, and core domain are colored blue, yellow, and pink, respectively.

3.1.1 Preparation for ensemble pool of GDH with continuous heterogeneity by molecular dynamics simulations. To evaluate the fitting performance of MDSPACE, we used synthetic cryo-EM 2D images derived from a known structure pool with continuous heterogeneity. This allows for a direct comparison between the ground truth (hereinafter referred to as the "target" pool) and fitted structures using the MDSPACE method. To construct such structure pools, we performed extensive all-atom MD simulations starting from the unligated state of GDH (Figure 3-1A) using the GENESIS MD simulation package. From the MD simulation trajectories, 2,000 frames are extracted to compose the structure pool for generating synthetic cryo-EM images. This pool shows high heterogeneity with a broad range of Cα-RMSD compared with the initial structure (PDB: 1EUZ) of sampling, including open, closed, and half-open states (representative models shown in Figure 3-2). This structure pool serves as a reliable ground truth for evaluating fitting performance. The trajectories that have been obtained will also be used later to examine experimental data.

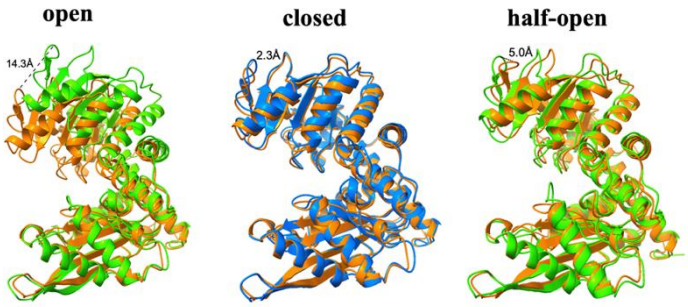

Figure 3-2: High flexibility of GDH is seen in the sampled structures in our preliminary MD simulation results. Three representative (open, closed, half-open) structures are shown.

3.1.2 Performance evaluation of the MDSPACE method with synthetic image data

2,000 synthetic cryo-EM images are generated from 2,000 structures sampled from MD simulation using the RELION package (Scheres, *J. Struct. Biol.* **180**, 519-530 (2012)). The images incorporate the contrast transfer function (CTF) with the parameters commonly found in experiments. The initial test was conducted with minimal background noise. Figure 3-3 illustrates a representative 3D structure and its cryo-EM projection images, which include CTF effect and noise.

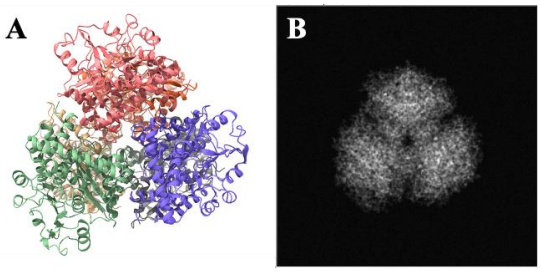

Figure 3-3: A representative of 3D structure (A) and its synthetic cryo-EM image (B).

MDSPACE analysis was then conducted using the aforementioned images as the targets. The initial conformation was derived from an X-ray structure of GDH, and 3D-to-2D flexible fitting calculations were performed through MD simulations with a modified version of GENESIS. The iterative conformational landscape refinement was performed with the NMMD integrator and Cα Go model using initially normal mode analysis and subsequently principal-component empowered MD simulations. This process gradually advances the conformations to align with the target images by applying a biased potential. As a result, each image is associated with a structure whose projection aligns with its synthetic image data.

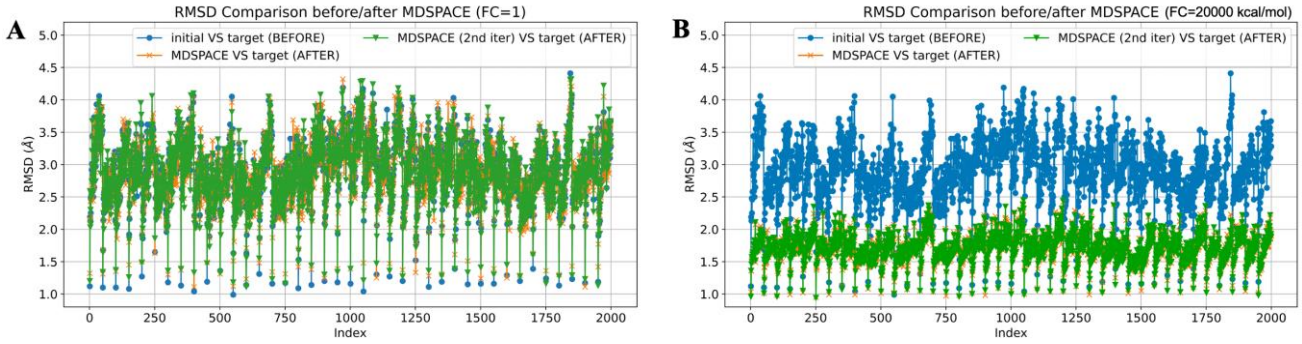

Figure 3-4: Cα-RMSD values as an indicator to evaluate the viability of MDSPACE method. The Cα-RMSD values are calculated by only considering the alpha carbon atoms of the protein. Two different force constant values are used for the biased potential: (A) force constant = 1 kcal/mol, (B) force constant = 20,000 kcal/mol.

A comparison of the results obtained with two different force constant values (Figure 3-4) reveals that the Cα-RMSD decreased by approximately 2 Å from the second iteration of the simulation and reached convergence for the remainder of the iterations when the force constant value was 20,000 kcal/mol. In contrast, the Cα-RMSD values remained relatively constant, essentially matching the starting point when the force constant was very small (1 kcal/mol). The comparison demonstrates the effectiveness of this method for identifying 3D structures matching the images.

3.1.3 Flexible fitting performance of different combinations of integrators and force fields

In the original approach, the flexible fitting protocol utilized the normal mode empowered MD simulations and Cα Go model. To obtain higher-resolution structure models from cryo-EM images and consider more motions than those

limited by the normal modes in MD simulations, more accurate models for proteins, such as the all-atom (AA) Go model, and other MD integrators, such as the LEAP and VVER integrator, were tested.

The MDSPACE method was evaluated in four scenarios: the NMMD integrator and Cα Go model, the NMMD integrator and AA Go model, the LEAP integrator and Cα Go model, and the VVER integrator and Cα Go model. Two indicators were used to assess the method's performance in each scenario. The Cα-RMSD values between the simulated and target structures and the cross-correlation coefficients (CC) between the target images and the simulated images projected from the simulated structures are shown in Figure 3-5. The AA Go model provides fitting results with all-atom information and comparable agreement to the ground truth as the Cα Go model, but its computation time is approximately 10 times longer.

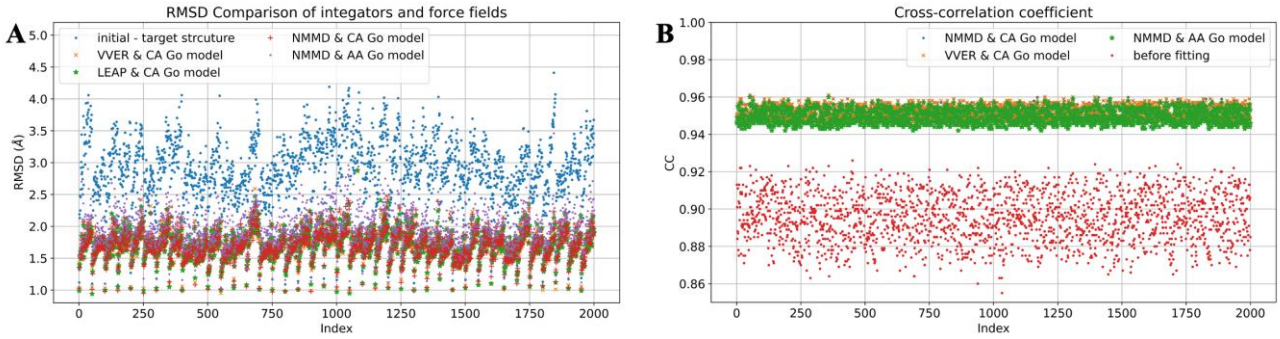

Figure 3-5: (A) Cα-RMSD values before (initial-target structure) and after MDSPACE flexible fitting (2000 images) with different integrators and force fields. (B) Cross-correlation coefficients between the target image and simulated image projected from the initial conformation (before fitting), and between the target image and simulated image projected from flexible fitted structures calculated with different combinations of integrators and force fields.

3.2 Optimize large-scale MDSPACE flexible fitting on Fugaku supercomputer

The original version of MDSPACE software uses a GUI protocol that is part of the Scipion package. The most time-consuming step is MDSPACE MD fitting (highlighted in red in Figures 3-6A). To take full advantage of the Fugaku supercomputer, we developed Shell and Python-based tools (Figure 3-6B) to enable the automatic execution of this step through a CLI.

In order to perform MD fitting on Fugaku for a large dataset, it is necessary to reorganize the input files, including the input EM image and parameter files. A hierarchical directory structure is created to avoid overloading the file system, which evenly distributes the input files among the directories. MD fitting simulations are then run in independent directories, enabling the handling of up to 100 million images (Figure 3-6D). Subsequently, our tool can set up multiple MD fitting simulations simultaneously as bulk jobs for maximum efficiency. Each job utilizes a single node to run multiple MD simulations simultaneously, for example, 12 simulations, and performs fitting via cycle over the targets (for example, 1000 times) for a large number of images, such as 12,000, on one node. Furthermore, a script has been created for post-simulation analysis, combining the MD simulation results and preparing the data for the next iteration. Finally, CLI-based Python scripts have been prepared to generate the outputs for the next protocol in

the project (Figure 3-6B).

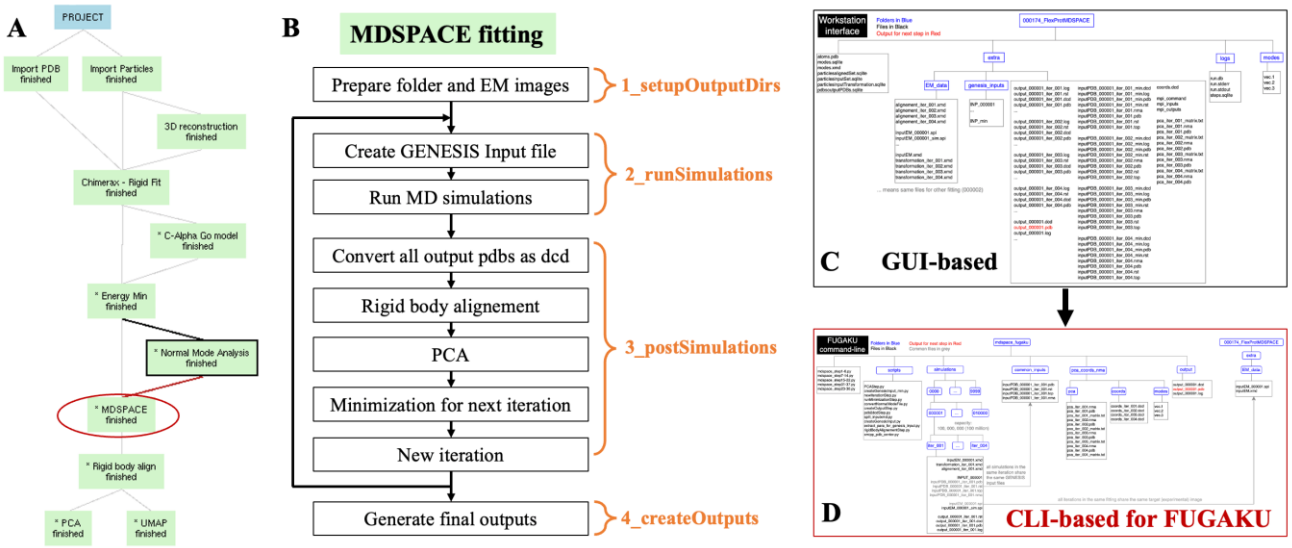

Figure 3-6: (A) Project structure of MDSPACE method in Scipion package. (B) Decomposition of MDSPACE protocol in GUI mode (text in black) and reformation in CLI mode (text in orange). (C) The default folder structure for the MD simulation step (MDSPACE, highlighted in red circle) in the project. (D) Newly designed folder structure for running iterative MDSPACE on Fugaku with a CLI-based tool compatible with Fugaku environment for large-scale flexible fitting MD simulation.

(サブ課題 4)エラー・ノイズを含む実験データからの統合的立体構造推論（東理大・森）

サブ課題 4 の主な成果として、クライオ電顕データとクロスリンク質量分析データから立体構造を MELD 法を用いて予測するための機能を MD 計算プログラム GENESIS に実装し、レプリカ交換法や全原子、粗視化モデルと組み合わせて立体構造をモデリングできるようにした。MELD 法は距離情報などの拘束条件に対して、実験データと合わないものを自動的に除外する方法であり、例えば、「実験データの 80%を利用する」ことを事前に指定して拘束しながら MD 計算を行う方法である。本研究で実装した MELD 法では、CHARMM および AMBER 力場に基づく全原子モデルや、郷モデルや AICG2+ model などの粗視化モデルが利用できるようにした。また、GB/SA モデルなどの陰的溶媒モデルにも対応させた。レプリカ交換法については、MELD 法で用いる拘束条件のエネルギー関数に対するスケーリングファクターの交換にも対応させた。

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ライオ電顕像データとクロスリンク質量分析データとを用いて電顕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。初期構造には、カルシウムイオンポンプの E2 状態 (PDB ID: 1IWO) を用い、ターゲットとするクライオ電顕マップは、E1·2Ca^{2+} 状態 (PDB ID: 1SU4) から解像度が 5 Å となるように生成した。また、クロスリンク質量分析データは、E1·2Ca^{2+} 状態の PDB 構造から生成し、タンパク質中の Lys-Lys ペアをランダムに選び、それを距離拘束条件とした。なお、本研究では、距離が 27 Å 以上のペアを overlength ペアとして定義し、このようなエラーを 20%クロスリンク情報に混入させた。なお、実験データ数依存性を調べるために、クロスリンク距離情報を 100, 200, 300 個とし、それぞれ 3 種

類のランダムなデータセットを用意した。本計算においては AICG2+粗視化モデルを用い、200 ps の電顕フレキシブルフィッティングを行った。MELD 拘束において利用するクロスリンク質量分析データを 0, 70, 80, 90%と変え、それぞれ 20 回の計算を実行した。計算の結果、クロスリンク質量分析データを用いない計算 (0%) すなわち、電顕データのみを用いて構造予測した場合には、マップが低解像であるために予測が成功したケースが 13/20 であったのに対して、クロスリンクデータを追加した場合、特に 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。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（図 4-1)。

より正確にエラー率依存性を調べるために、MELD 拘束において利用するクロスリンク質量分析データを 60%から 100%の間において 2%刻みで変えて構造を予測した（図 4-2)。実験データには 80%の真のデータによって構成されているが、実際には 88%のデータを利用しても、すなわち 8%分のエラーを含めても、20 回中 20 回ほぼ正しい構造が得られることがわかり、正しい構造を予測するためには、特に 27 Å 付近の構造情報が豊富に含まれることが重要であることが分かった。以上の結果から、クロスリンク質量分析データ中のエラー率がおおよそわかれば、MELD 法において実験値と一致しないデータをエラーとみなして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。

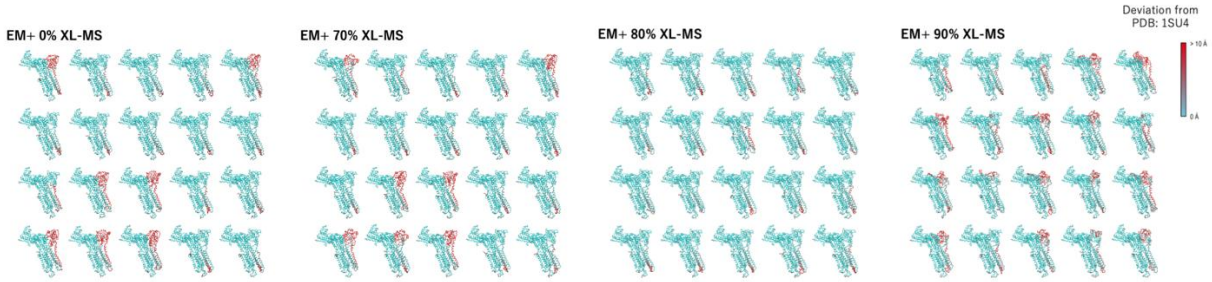

図 4-1: 0, 70, 80, 90%のクロスリンク質量分析データを加えたフレキシブルフィッティング計算の結果

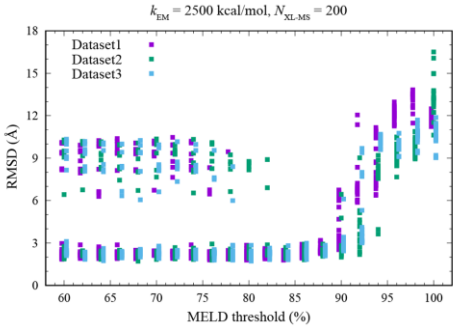

図 4-2: 20%のエラーを含むクロスリンク質量分析データの利用を 60%から 100%まで変化させた時の最終構造と正解構造との RMSD の比較

本研究課題では、東工大・野澤らによって得られたヌクレオソーム複合体に対するクロスリンク質量分析実験データに対しても解析を行った。クロスリンク距離を正確に定義するために、まず、実験で用いられたクリスリンカーDSS（disuccinimidyl suberate ; 図 4-3A）の両端に Lys を結合させた分子に対して MD 計算を行い、クロスリンク距離を精密に求めた。次に、複合体に対する MD 計算を行い、複合体内のクロスリンクペア距離を解析した。複合体構造 (原子数：1,025,464 個) を 150 mM KCl 溶液に浸し、298.15

K、1 atm の条件下で 200 ns の MD 計算を実行した。計算には MD 計算ソフトウェア GENESIS を用いた。DSS に結合した Lys の Cα 原子間距離を解析した結果、DSS は 20 ns の間大きく構造変化し、距離の最大値は、24.8 Å にもなることがわかった。このことから、DSS はタンパク質内において 25 Å 以内の Cα 原子間距離を持つ Lys を架橋する可能性が示唆された。図 4-3B に、実際のクロスリンク質量分析実験で架橋された Lys-Lys ペアの MD トラジェクトリー中での Cα 原子間距離の解析結果を示す。クロスリンク質量分析実験データの 68.9 %程度は、本来架橋しないはずの 25 Å 以上離れた Lys-Lys ペアであることが分かり、これらのデータはタンパク質の会合や凝集に由来するものであると考えられた。今後、このような情報に基づいて、MELD 法を用いた立体構造予測を進める予定である。

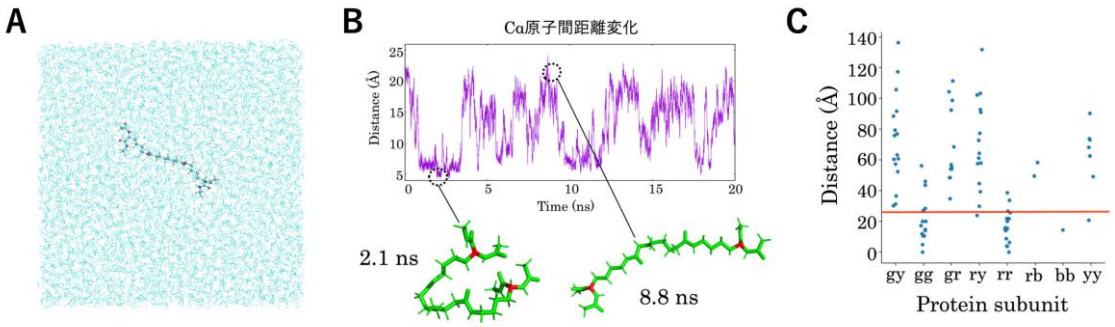

図 4-3: (A) リシンを両端に結合させたクロスリンカーDSS, (B) DSS の Cα 原子間距離の時間変化, (C) ヌクレオソーム複合体のクロスリンクペアの MD トラジェクトリー中の距離

(GENESIS への機能実装)

上記のサブ課題で実装された実験データからの大規模推論のために MD ソフトウェア GENESIS へ実装した機能は以下の通りである。

- 「富岳」上で多量の CryoEM 画像データについて並列的に GENESIS による 2D フレクシブルフィッティングを行う処理スキームを開発した(MDSPACE アルゴリズムへの対応)。
- クライオ電顕データとクロスリンク質量分析データへ応用可能な MELD 法を GENESIS へ実装し、CHARMM および AMBER 力場に基づく全原子モデルや、郷モデルや AICG2+ model などの粗視化モデルが利用できるようにした。また、GB/SA モデルなどの陰的溶媒モデルにも対応させた。
- GENESIS へ既実装のレプリカ交換法を、MELD 法で用いる拘束条件のエネルギー関数に対するスケーリングファクターの交換にも対応させた。

(プロジェクトの総合推進)

プロジェクト全体の連携を密としつつ円滑に運営していくためのミーティングを、Zoom で月に 1 回開催した。課題の全員が参加し毎月 1 チームずつが順番に進捗報告を行い、その後に運営のための短い PI ミーティングを行った。また、2024 年 1 月に課題内ワークショップを対面で開催し、若手（研究員・学生）も含めて進捗報告を行うとともに、GENESIS と機械学習/AI の連携について密に議論した。日頃の運営に必要なコミュニケーションは Slack で行い、ファイル共有は OneDrive で行った。

２－３．活動（研究会の活動等）

国際ワークショップ "Multi-scale Molecular Dynamics Simulation and Machine Learning of Biomolecular Systems" の開催

2023 年 8 月 9 日-10 日に国際ワークショップ "Multi-scale Molecular Dynamics Simulation and Machine Learning of Biomolecular Systems" を理化学研究所和光キャンパスにおいて、杉田理論分子科学研究室(杉田有治主任研究員、八木清専任研究員)、名古屋大学岡本祐幸教授、と本課題が共催して開催した。ワークショップでは、今後の分子シミュレーションと機械学習/AI との連携について活発に議論した。対面のみで開催し 48 名が参加した。2 日目に本課題のセッションを設けて、本課題から 4 名が招待講演を行い、本課題を宣伝するとともに、これまでに得られた成果について報告した。招待講演者 (下線は海外の招待講演者)は以下の通りである。

- Giovanni Brandani (Kyoto University, Japan)
- Bernard Brooks (NIH, USA)
- Fumio Hirata (IMS, Japan)
- Jaewoon Jung (RIKEN R-CCS, Japan)
- Akio Kitao (Tokyo Institute of Technology, Japan)
- Juyong Lee (Seoul National University, Korea)
- Mai Suan Li (Polish Academy of Science, Poland)
- Yasuhiro Matsunaga (Saitama University, Japan)
- Osamu Miyashita (RIKEN R-CCS, Japan)
- Toshifumi Mori (Kyushu University, Japan)
- Ai Niitsu (RIKEN CPR, Japan)
- John Straub (Boston University, USA)
- Cheng Tan (RIKEN R-CCS, Japan)
- Kiyoshi Yagi (RIKEN CPR, Japan)

課題内ワークショップの開催

2024 年 1 月 9 日-10 日に課題内ワークショップを理化学研究所神戸キャンパスにおいて対面とリモートで開催した。まだ非公開の成果を含めた今後の研究について深く議論するためにクローズドな形で開催した。1 日目に各グループからの成果報告を行うとともに、理化学研究所計算科学研究センターの杉田有治チームリーダによる招待講演を設けた。2 日目は、GENESIS と機械学習/AI の連携について議論する場を設けるとともに、若手育成のために研究員・学生のショートトークセッションを設けた。

マンスリーセミナーの開催

2023 年中は課題内での月一回の進捗報告をオンライン(Zoom)で行っていたが、2024 年 1 月から外部の招待講演者を迎えて分子シミュレーションと機械学習/AI に関する Zoom セミナーを開催した。1 月は名古屋大学佐久間航也博士による AlphaFold 解説、2 月は理化学研究所徳久淳師博士による「富岳」での OpenFold 実装に関するセミナーを行って頂いた。次年度も引き続き月一回で開催し、AI との連携に関する分野の底上げに取り組んでいきたい。

２－４．実施体制

業務項目	担当機関	担当責任者
（１）（サブ課題 1）マルコフ状態モデルを介した構造ダイナミクスの推論	〒338-8570 埼玉県さいたま市桜区下大久保 255　国立大学法人埼玉大学	松永 康佑
（２）（サブ課題 2）遺伝子発現と共役するゲノム 3 次元構造動態の推論	〒606-8501 京都府京都市左京区吉田本町 36 番地 1　国立大学法人京都大学	高田 彰二
（３）（サブ課題 3）クライオ電顕データからの構造アンサンブル推論	〒464-8601 名古屋市千種区不老町 国立大学法人東海国立大学機構 名古屋大学	Florence Tama
（４）エラー・ノイズを含む実験データからの統合的立体構造推論	〒162-8601 東京都新宿区神楽坂1-3　学校法人 東京理科大学	森 貴治
(プロジェクトの総合推進)	〒338-8570 埼玉県さいたま市桜区下大久保 255　国立大学法人埼玉大学	松永 康佑

別添１　学会等発表実績

１．学会誌・雑誌等における論文掲載

No.	掲載した論文（発表題目）	発表者氏名	発表した場所（学会誌・雑誌名等）	発表した時期
1	Micelle-like clusters in phase-separated Nanog condensates: A molecular simulation study.	Azuki Mizutani(Kyoto University) Cheng Tan(RIKEN) Yuji Sugita(RIKEN) Shoji Takada(Kyoto University)	PLOS Computational Biology, , 19(7): e1011321 (2023)	2023/07
2	Postsynaptic protein assembly in three and two dimensions studied by mesoscopic simulations.	Risa Yamada(Kyoto University) Shoji Takada(Kyoto University)	Biophysical Journal, Vol.122, pp.3395-3410 (2023)	2023/08
3	Representation of Protein Dynamics Disentangled by Time-Structure-Based Prior.	Tsuyoshi Ishizone(Meiji University) Yasuhiro Matsunaga(Saitama University) Sotaro Fuchigami(University of Shizuoka) Kazuyuki Nakamura(Meiji University)	J. Chem. Theory Comput. Vol. 20, pp. 436–450 (2024)	2023/12

２．国際会議・シンポジウムにおける口頭・ポスター発表

No.	発表した成果（発表題目、口頭・ポスター発表の別）	発表者氏名（所属機関）	発表した場所（学会名等）	発表した時期
1	Analysis of structural changes of multi-chain/multi-domain proteins by coarse-grained description. (ポスター)	Chigusa Kobayashi(RIKEN R-CCS), Hisham M. Dokainish(RIKEN CPR), Suyong Re(NIBIOHN), Takaharu Mori(RIKEN CPR), Jaewoon Jung(RIKEN R-CCS, RIKEN CPR), Yuji Sugita(RIKEN R-CCS, RIKEN CPR, RIKEN BDR)	CCP2023 - 34th IUPAP Conference on Computational Physics	2023/08
2	Accurate Langevin integration for isothermal-isobaric condition with a large time step. (口頭)	Jaewoon Jung (RIKEN R-CCS, RIKEN CPR) Yuji Sugita (RIKEN R-CCS, RIKEN CPR, RIKEN BDR)	CCP2023 - 34th IUPAP Conference on Computational Physics	2023/08

3	Integrative Modeling of Biomolecular Dynamics from Molecular Dynamics Simulations and Single-Molecule Experiments. (口頭, 招待講演)	Yasuhiro Matsunaga (Saitama University)	Multi-scale Molecular Dynamics Simulation and Machine Learning of Biomolecular Systems	2023/08
4	Development of GENESIS on Fugaku for Large-Scale MD Simulations. (口頭, 招待講演)	Jaewoon Jung (RIKEN R-CCS)	Multi-scale Molecular Dynamics Simulation and Machine Learning of Biomolecular Systems	2023/08
5	Integrating Hi-C Data into Polymer Simulations to Study the Principles of Enhancer-Promoter Communication. (口頭, 招待講演)	Giovanni Brandani (Kyoto University)	Multi-scale Molecular Dynamics Simulation and Machine Learning of Biomolecular Systems	2023/08
6	Integrative/Hybrid Modeling Approaches for Dynamic Structural Biology. (口頭, 招待講演)	Osamu Miyashita (RIKEN R-CCS)	Multi-scale Molecular Dynamics Simulation and Machine Learning of Biomolecular Systems	2023/08
7	Integrative modeling of biomolecular dynamics from molecular dynamics simulations and single-molecule experiments. (口頭, 招待講演)	Yasuhiro Matsunaga (Saitama University)	The 6th International Conference on Molecular Simulation (ICMS2023)	2023/10
8	Coarse-grained description of structural changes for multi-	Chigusa Kobayashi (RIKEN R-CCS) Yuji Sugita (RIKEN R-CCS, RIKEN CPR, RIKEN BDR)	The 6th R-CCS Symposium	2024/01

	chain/multi-domain proteins. (ポスター)			
9	Development of coarse-grained MD program for large-scale simulations of heterogeneous biomolecules. (ポスター)	Jaewoon Jung (RIKEN R-CCS, RIKEN CPR) Cheng Tan (RIKEN R-CCS) Yuji Sugita (RIKEN R-CCS, RIKEN CPR, RIKEN BDR)	Biology of Intracellular Environments International Symposium 2024	2024/02
10	Coarse-grained methods for proteins with multi-chain/multi-domain (ポスター)	Chigusa Kobayashi(RIKEN R-CCS) Hisham M. Dokainish(RIKEN CPR) Suyong Re(NIBIOHN) Takaharu Mori(RIKEN CPR) Jaewoon Jung(RIKEN R-CCS, RIKEN CPR) Yuji Sugita(RIKEN R-CCS, RIKEN CPR, RIKEN BDR)	Biology of Intracellular Environments International Symposium 2024	2024/02