2020 年度　利用報告書

課題名(タイトル):分子動力学シミュレーション解析人工知能システムの応用研究

利用者氏名:○山根　努、三十尾　潔高
理研における所属研究室名:医科学イノベーションハブ推進プログラム　医薬プロセス最適化プラットフォーム推進グループ　分子設計インテリジェンスユニット

1.　本課題の研究の背景、目的、関係するプロジェクトとの関係

タンパク質の分子動力学(MD)シミュレーションでは非常に多くのダイナミクスの経時変化のデータ(トラジェクトリ)が生成される。しかし、その情報量は膨大であり、その中からタンパク質の構造変化や運動にかかわる重要な特徴(例:特定の原子間の距離、タンパク質の主鎖の二面角等)を導き出すのは、非常に困難である。

このような背景から、医科学イノベーションハブプログラムにおける、ライフインテリジェンスコンソーシアム(LINC)において"分子動力学シミュレーション解析人工知能システム"(gini ツール)を開発した。このシステムでは分子動力学シミュレーションから得られたトラジェクトリからその構造変化やダイナミクスを表現することのできる特徴量(フィーチャー)を機械学習により得られた gimi importance と呼ばれる指標により、検出することができる。

本課題では、このシステムをいくつかの分子動力学シミュレーションの系に適用することで、このシステムの応用と高度化を目指した。また、MD シミュレーションの解析手法として近年多く用いられてきているマルコフ状態モデル(MSM)解析法においても、重要な特徴量の抽出を行うことが一つの問題課題となっている。そこで、MSM を利用する解析手法である MD-SAXS 法の開発を行った。

2.　具体的な利用内容、計算方法

本課題での具体的な利用方法は以下の 2 つである。

(1) gini ツールの検証

開発された gini ツールが、仕様通りの挙動を示し、MD シミュレーションのトラジェクトリを分類しうる特徴量を検出することができるかの検証を行った。計算には、環状ペプチドであるシクロスポリン A の異なる 2 つの構造(オープン、クローズド、Figure 1)を用いた水中での 1 μ 秒の MD により得られたトラジェクトリを用いた。ここでは、シクロスポリン A 分子の主鎖骨格の 3 種類の二面角(φ, ψ, ω)およびアミノ酸の α 炭素間の距離(Figure 2)を特徴量として選んだ。

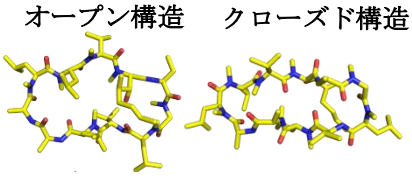

Figure 1:シクロスポリン A の 2 つの構造状態

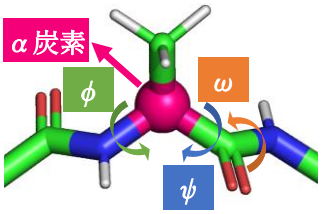

Figure 2:アミノ酸の α 炭素および二面角(φ, ψ, ω)

(2) MD-SAXS システムの開発

上述の gini ツールにより検出されるような、特徴量をどのように探索するのかは、現在分子動力学シミュレーションのトラジェクトリの解析で多く用いられている手法であるマルコフ状態モデル(MSM)解析においても、共通の問題である。したがって、本課題では MSM 解析を用いた MD-SAXS システムの開発をおこなった。

MSM 解析では、目的のタンパク質の MD シミュレーションにより得られた多数のトラジェクトリを用い、その中に含まれる構造を分類し、多数の構造の集まり(クラスタ)を作成する。ここで得られたクラスタをもとに、クラスタの間でどのくらいの頻度で構造の変化が起きているのかを解析し、それぞれのクラスタに属する構造の生じやすさ(確率重み)を求める。この解析により目的のタンパク質が水中でどのような構造の状態の間をどのくらいの頻度で変遷しているのかの情報を得ることができる(Figure 3)。このとき、クラスタを分類する際に、構造の違いをうまく表現することのできる特徴量を見つ

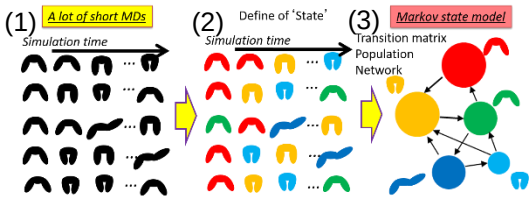

Figure 3:マルコフ状態モデルの説明
(1) 多数の MD トラジェクトリを集める。
(2) 得られた構造を分類する。
(3) 各構造状態の存在確率および状態間の変化の確率を調べる。

けることが、解析をうまく行うカギとなる。

一方で、タンパク質の溶液中での構造の解析を行う手段の一つとして、小角 X 線散乱(SAXS)という方法がある。この解析により、散乱プロファイルと呼ばれる観測値が得られる。これは、タンパク質中の水素原子以外の原子間距離の頻度分布が与えられ、タンパク質分子の外形を反映したものである。タンパク質の一つの構造に対して一つのプロファイルが与えられるが、水溶液中ではタンパク質が様々な構造状態の間を変化しており、SAXS プロファイルの観測値は、それらの構造から得られる統計平均と考えられる。したがって、MD シミュレーションの結果の MSM 解析で得られた構造状態の確率重みを用いれば、より厳密な、水溶液中での散乱プロファイルを再現することができる(Figure 4)。

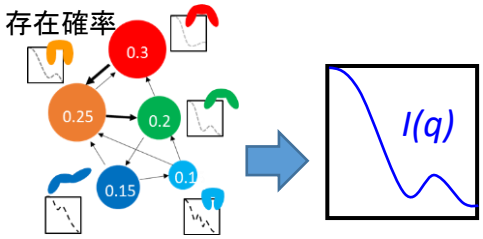

Figure 4: MSM の SAXS プロファイルへの適用

ここでは、SAXS プロファイルを再現するような構造アンサンブルのトラジェクトリを PaCS-MD と呼ばれる MD 計算の手法から求める方法(PaCS-Fit)および、得られたトラジェクトリから、MSM 解析を用いて、溶液中での SAXS プロファイルを正

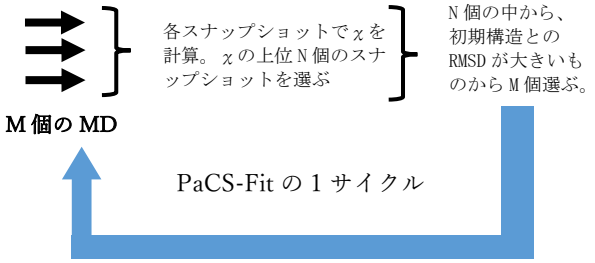

Figure5:PaCS-Fit 法の説明

確に再現する方法の開発を行った。PaCS-Fit 法は、短い MD シミュレーションを同時に M 本流し、そのシミュレーションの途中のすべての構造の中で、ターゲットの SAXS プロファイルに対する、類似性の指標である χ 値を計算し、χ 値に基づき最もターゲットの SAXS プロファイルに近く、かつ初期構造との RMSD が大きいものを M 個選び、それらを初期構造としてまた MD を N 本流す(Figure 5)。プロセスを繰り返すことで、ターゲットの SAXS プロファイルを再現する構造に収束させた構造を集める。本研究では、M, N の値はそれぞれ、10, 20 であり、各 MD シミュレーションは 100ps 実行し、プロセスは 50 回繰り返した。

計算には、オープン構造とクローズド構造の 2 つの構造が知られているアデニル酸キナーゼ(ADK)を用いて以下の手順で行った。

1. オープン構造の水中での MD を 1 μ 秒 MD プログラム GROMACS により行い、得られたトラジェクトリのすべての構造について、SAXS の理論散乱プロファイルをプログラム CRYSOL にて求めた。
2. 1. で求めたオープン構造の水中での理論散乱プロファイルをターゲットとして、クローズド構造を初期構造とした PaCS-Fit を実行し、オープン構造の散乱プロファイルを再現する構造のサンプリングを実施した。
3. 2. で求めたクローズド構造からオープン構造への変化のトラジェクトリを用いて、MSM 解析を実施し、クラスタリングにより多数の構造状態に分離し、各状態確率密度を求める。この際に、構造状態を分離する特徴量として、クローズド構造とオープン構造の違いをうまく分離しうると考えられる、アデニル酸キナーゼのアミノ酸の α 炭素の原子間距離の 958 組を用い、tICA 法を用いて、構造を分利することができる上位 5 つの特徴量を探索して、もちいた。さらに、これらの各構造状態に対し SAXS プロファイルを求め、各状態の確率密度による重み平均から、SAXS プロファイルを算出する。得られた SAXS プロファイルとターゲットの SAXS プロファイルを比較し、PaCS-Fit により、オープン構造がサンプリングできたかの確認を行う。ここで、MSM 解析はプログラム MSMBuilder を用いた。

3. 結果

(1) gini ツールの検証

シクロスポリン A のオープン構造およびクローズド構造を初期構造した、水中で各 1 μ 秒の MD シミュレーションで得られたトラジェクトリを用い、これらのトラジェクトリの構造を分類することのできる特徴量を検出することができるかの、gini ツールの検証を行った。これらのトラジェクトリについて、主鎖の 3 つの二面角(ϕ, ψ, ω)の sin, cos での値および α 炭素間距離の 4 つの特徴量について gini ツールを実行し、構造をうまく分類する特徴量を検出できるのかを検証した。、機械学習は、ランダムフォレストを用いて行った。gini importance の上位5個の特徴量を示す(Table 1)。

Table 1: gini importance の高い上位 5 つの特徴量

ID	Gini importance	Feature type	Position
1	0.0857	psi_sin	MLE8
2	0.0849	Distance CA	MLE3-VAL9
3	0.0846	Distance CA	MLE2-VAL9
4	0.0842	Distance CA	DALA1-MVA4
5	0.0839	omega_cos	MLE2-MLE3

その結果、2 つのアミノ酸 MLE2, MLE3（MLE:N-メチルロイシン）の 2 面角が gini importance の上位を占めていた。これらの上位 2 つの特徴量でのトラジェクトリ中の構造の分布のプロットを示す（Figure 6）。

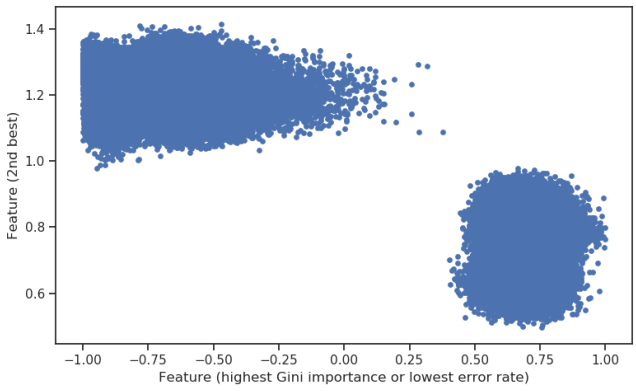

Figure 6:gini importance 上位 2 つの特徴量によるトラジェクトリ構造の分布

この図に示されるように、トラジェクトリ中の構造がきれいに 2 つに分かれて分布しており、得られた特徴量が構造の特徴を分類できることが示唆された。得られた上位5つの特徴量をシクロスポリン A のオープン構造とクローズド構造の上にマッピングした図を示す（Figure 7）。

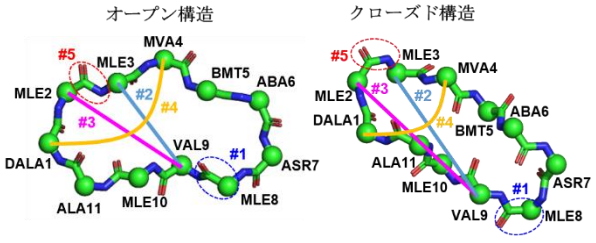

Figure7:gini importance 上位 5 つの特徴量のシクロスポリン A 上の位置

ここでもっとも大きな gini importance を示した、MLE8 の ϕ は、オープン構造とクローズド構造でそれぞれ、103.5° および 32.8° と変化量が大きく、これらの構造全体の形状変化にかかわっている可能性があることが分かった。また、その次に大きな gini importance を示した 3 つの α 炭素間距離、MLE2-VAL9, MLE3-VAL9 および DALA1-MVA4 は、Figure6 に示すように、オープン構造とクローズド構造の間で距離が大きく変化していることがわかる。さらに、シクロスポリン A のオープン型とクローズド型の構造の間では、MLE2-MLE3 の間の二面角 ω がそれぞれ、トランス型（180°）、シス型（0°）を取ることが知られており（Figure 6 の赤点線丸で示した部分）、そのこと違いを検出することができたと考えられる。

(2) MD-SAXS システムの開発

アデニル酸キナーゼ（ADK）のオープン構造を初期構造として、1 μ s の MD シミュレーションを実行し、得られたトラジェクトリの全構造から求めた、理論 SAXS プロファイルの平均値を算出した。この SAXS プロファイルをターゲットとし、クローズド構造を初期構造として、PaCS-Fit 法を実行した。PaCS-Fit 法の計算は HOKUSAI

BW により実行した。PaCS-Fit 法のプロセスの繰り返しにおいて、各プロセスでの構造がどの程度ターゲットの SAXS プロファイルに近づいたかを示す指標（χ 値）の変化を示す（Figure8）。

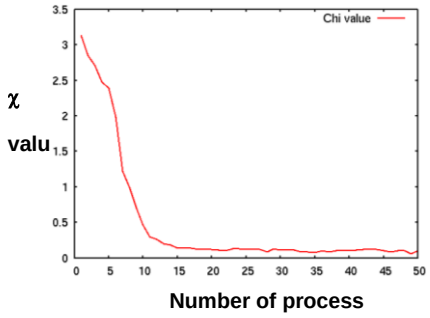

Figure 8:PaCS-Fit 法における、プロセス毎の χ 値の変化

この図からわかるように、50 回のプロセスを通して、χ 値は 3.13 から 0.10 まで減少しており、このことは、SAXS プロファイルがよりターゲットの値に近づいたことを示している。ここで得られた、全 50 プロセスのトラジェクトリ 500 本（各 100ps）に対して MSM 解析を行い、構造状態をよりよく分離する特徴量を tICA 法により求めた。上位の 2 つの特徴量で表される平面上に投影した図を示す（Figure 9）。

図より、トラジェクトリ中の全構造は、2つの特徴量の平面上で扇状の帯となって分布しており、それぞれの場所での構造が、オープンからクローズドまでの構造の変化と対応している。

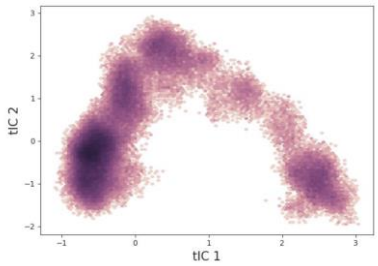

Figure9:PaCS-Fit により求めた、トラジェクトリ構造の、tICA 法により求めた特徴量上位 2 つによる平面上へのプロット

また、構造をクラスタリングし、MSMにより得られた分布の様子を示す(Figure 10)。

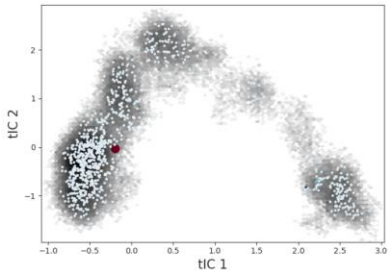

Figure10:MSM 解析で得られた各クラスタ中心の分布（図中の白丸のプロット）

Figure 10 より、MSM 解析では tICA により得られた平面上の各位置に分布するクラスターをくまなく用いて、クラスター間の確率過程を表現することができていることが示唆される。

ここで得られた、各点の構造に対して、理論SAXSプロファイルを求め、ターゲットのプロファイルと比較した(Figure 11)

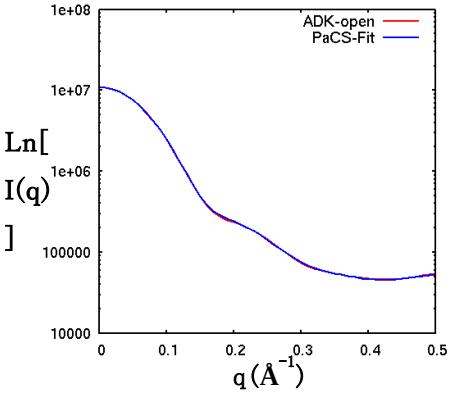

Figure11:ターゲットの SAXS プロファイル(赤) と PaCS-Fit 法により得られた SAXS プロファイル（青）の比較

この図より、ここで開発した方法で得られた、理論SAXSプロファイルが、ターゲットのSAXSプロファイルとよく一致することが明らかとなった。

4.　まとめ

本課題ではMDシミュレーションの解析を行う上で重要な特徴量を扱う2つの研究として、(1)gini ツールの検証および(2)MD-SAXS 法の開発をおこなった。(1)の gini ツールの検証は、シクロスポリン A の異なる初期構造であるオープン構造およびクローズド構造のからの MD シミュレーションの結果を分離する特徴量をうまく検出することができた。シクロスポリン A は親水溶媒、疎水溶媒中でそれぞれオープン構造、クローズド構造が安定であり、この構造間の転移が膜透過などの機能に重要であると考えられている。今回検出することのできた、これらの特徴量およびその周辺の構造に関する解析を行うことで、この構造変化のメカニズムを理解することができると考えられる。

(2)の MD-SAXS 法の開発では、PaCS-Fit 法という新規な MD シミュレーション法の開発を行い、これによるターゲットの SAXS プロファイルを再現する構造アンサンブルを得ることが可能であることが示された。また、この過程で用いた MSM 解析では、異なる構造状態を分離し、ターゲットの SAXS プロファイルを再現することができた。

5.　今後の計画・展望

giniツールの検証においては、今回の解析ではシクロスポリン A の構造にかかわる特徴量(原子間距離および二面角)を用いた。しかし、シクロスポリン A のオープン構造とクローズド構造では、安定に存在する溶媒環境が異なることから、周辺溶媒との相互作用の違いも特徴量として用いることができると期待される。そこで、giniツールでタンパク質溶媒間の相互作用を特徴量として用いることができるように、現在改良中であり、改良後にこれらの方法での検証を行う予定である。

また、MD-SAXS法については、今回はMSM解析において、tICA 法を用いて、特徴量を求めた。今後、この特徴量から MSM 解析により得られた異なる構造状態が、本研究で用いたアデニル酸キナーゼのオープン構造とクローズド構造の間をつないだ確率過程の検証を行う。また、gini ツールにより、検出された特徴量を用いて、MSM 解析を検証する予定である。