Vol.2013-MPS-96 No.7
Vol.2013-BIO-36 No.7
2013/12/11

グラフを用いたタンパク質のモーション表現法の提案

安富祖　仁[†1]　峯田克彦[†2]　遠藤俊徳[†2]

本研究では、分子動力学法 (Molecular Dynamics, MD) を用いたシミュレーションによって得られるタンパク質の時系列座標データ (トラジェクトリ) からの非線形モーション抽出手法を提案した。従来、トラジェクトリからのモーション抽出には主成分分析 (Principal Component Analysis, PCA) を利用した Essential Dynamics Analysis などの手法が用いられてきたが、線形手法に基づく方法ではタンパク質の非線形モーションの扱いが難しいという問題があった。そこで本研究では、タンパク質構造のクラスタリングと構造間の時間的隣接関係を利用して、トラジェクトリ内の構造遷移を表現する有向グラフ (構造遷移グラフ) を構築し、構成されたグラフ内の各辺に対応する構造変化を可視化することで、タンパク質の非線形モーションを抽出する手法を提案した。提案手法を Villin headpiece subdomain(HP-35 NleNle) のトラジェクトリに対して適用した結果、タンパク質の折り畳み過程において生じている部分的回転という非線形モーションの抽出が可能であることが示された。また、構造遷移グラフの形状が従来研究の結果に対応していることも確認された。

1. 背　　景

計算機ハードウェアの分野において、スーパーコンピュータ京の開発など、計算機リソースの飛躍的な性能向上が実現されている。またアルゴリズムの分野においては、タンパク質の平衡状態における相互作用をバネとして表現することによって、構造的揺らぎ (Fluctuation) の近似的記述を行う粗視化手法 1 つである Elastic Network Model[1] など、様々な効率的なアルゴリズムが開発されてきた。これらの計算機ハードウェアとアルゴリズム面の両面における発展により、タンパク質がペプチド鎖から折り畳まれて形成される折り畳み (Folding) と呼ばれる過程や、受容体 (Receptor) というあるタンパク質に、別のリガンド (Ligand) と呼ばれるタンパク質が結合するリガンド結合 (Ligand Binding) という過程などを計算機上でシミュレーションすることが可能となった。[2][3]。さらにモデル構築という観点からは、ある細胞の表現型 (Phenotype) を遺伝子型 (Genotype) から予測するモデルも提案されている[4]。生物学的プロセスの計算機上でのシミュレーションにより、生物学的実験から得られた知見のメカニズムの予測や、逆にシミュレーションから得られた知見を用いた新たな生物学的実験のデザインが可能となり、これにより生物学的プロセスに関する詳細な知識が得られると考えられている。生物学的プロセスのメカニズムの詳細な理解によって、アルツハイマー病 (Alzheimer's Disease) をはじめとする疾患の治療などの応用が実現されると期待されている[5]。

計算機上でタンパク質のような生体内分子の挙動をシミュレーションするために、分子動力学法 (Molecular Dynamics、MD)[6] が用いられてきた。MD シミュレーションを行うための代表的なソフトウェアとして、AMBER[7] や GROMACS[8] などがあげられる。また、効率的にタンパク質の構造空間 (Conformation 空間) をサンプリングするために、レプリカ交換分子動力学法 (Replica Exchanging Molecular Dynamics, REMD)[9] といった手法が提案されてきた。さらに、分散コンピューテイングの枠組みを利用してタンパク質折り畳みのシミュレーションを行う Folding@home[10] というプロジェクトも遂行されてきた。

2. 従来法とその問題点

MD から得られるタンパク質、あるいは着目する系全体の原子の時系列座標データをトラジェクトリ (Trajectory) と呼ぶ。トラジェクトリから着目するタンパク質のトラジェクトリには、タンパク質の機能に関連する原子のモーションだけでなく、系の熱的な揺らぎ (Thermal Fluctuation) によって生じるランダムな座標変動も含まれる。そこで、タンパク質の機能のメカニズムを理解するためには、トラジェクトリからタンパク質の機能に関連していると考えられる本質的なモーションを抽出する必要がある。そのための代表的な手法として、主成分分析 (Principal Component Analysis, PCA) を利用した Essential Dynamics Analysis(EDA)[11] が挙げられる。EDA では、まずタンパク質を構成する各原子のトラジェクトリにおける平均位置からの座標変動 (Root Mean Square Fluctuation,RMSF) の共分散行列 Σ を求める。次に共分散行列 Σ の固有値 λ_i と対応する固有ベクトル v_i を求める。

†1 北海道大学知識メディアラボラトリ
†2 北海道大学大学院情報科学研究科

Vol.2013-MPS-96 No.7
Vol.2013-BIO-36 No.7
2013/12/11

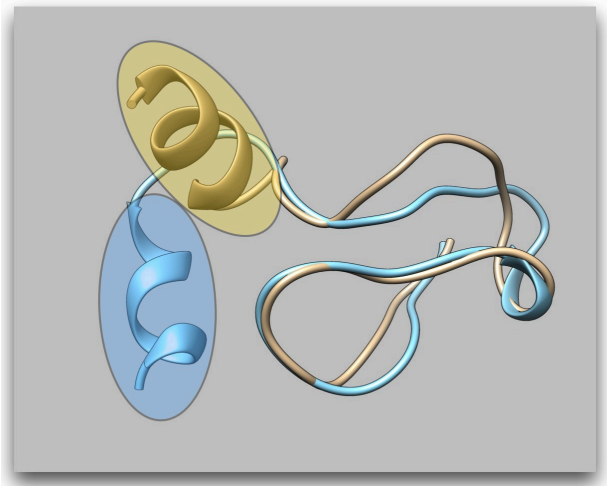

図 1 EDA では扱いにくいタンパク質モーションの例: 図中の右側の部分は比較的構造がマッチしているが左側の部分に回転運動が生じている。

求められた固有ベクトルのうち、絶対値の大きな固有値に対応する固有ベクトルはトラジェクトリに含まれる周波数の低いモーションに対応している。一方で、タンパク質の熱的揺らぎによる運動は高い周波数を持つ。そこで、トラジェクトリの各時点におけるタンパク質の構造を表現するベクトル (タンパク質内の各原子の座標を並べたベクトル) を固有ベクトル上に射影することによって、ある時点における熱的揺らぎを取り除いた本質的なモーションを抽出することが可能となる。以上が EDA のモーション抽出ステップである。

EDA によるモーション抽出は、タンパク質のトラジェクトリにおける長期的なモーションを抽出するために広く用いられてきた。しかしそのような成功の一方で、EDA には実用上の問題点が存在することも指摘されてきた。それは、EDA が線形手法である PCA を利用した手法であるため、ある短い時間内に生じるタンパク質の部分的な回転などの非線形モーションを扱うことが難しいという点である。図 1 に EDA では扱いにくいタンパク質モーションの例を示す。 図 1 において、網掛け円で示された部分が回転運動によって変形している部分を示している。部分回転のような非線形モーションは、α-helix や β-hairpin などのタンパク質の機能に関連する 2 次構造の形成に関連していると考えられるため、これらのモーションがトラジェクトリのどの時点において、どのように生じているかを理解することで、タンパク質の機能に関連した構造に関する知識が得られると考えられるが、線形手法である EDA ではこのようなモーションを抽出することが難しい。

そこで Isomap をはじめとする多様体学習に基づくタンパク質の非線形モーション抽出手法が提案されてきた[12)]。多様体学習に基づく手法では、データ分布の局所的な情報から、サンプル間の大域的な距離情報を復元し、その距離を用いて一般に高次元ベクトルで表現されるタンパク質構造をより低次元の空間で表現する。低次元で表現されたタンパク質構造の分布をクラスタリングなどの手法によって分類し、クラスタ間の違いを調べることにより、タンパク質の非線形なモーションを抽出できるという利点がある。しかし、多様体学習に基づく方法では与えられたデータが well-sampled、すなわち特徴空間上を十分な密度でサンプリングして得られたことを仮定しているため、データを表現する多様体内に点密度が周囲と比較して過度に低い領域が存在する場合には多様体が分割され、低次元表現が適切に行われないという問題がある。このような多様体における点密度の変化は、大規模なトラジェクトリのサブサンプリングやタンパク質構造の急激な変化によって引き起こされるため、このような処理や変化が含まれるデータセットを扱う際には単純な多様体学習法ではモーション抽出が困難となる。

3. 状態遷移グラフに基づくタンパク質モーションの抽出手法

タンパク質の非線形モーション抽出のため、本研究では構造クラスタリングとトラジェクトリ内の時間的隣接関係を利用した状態遷移グラフに基づくモーション抽出手法の提案を行った。図 2 に提案手法の概要を示す。 本節ではまず、構造クラスタリングの必要性について説明する。前述のように、MD シミュレーションから得られるトラジェクトリには、熱的揺らぎによって生じるタンパク質の機能には直接関連のない座標変動も含まれる。このような座標変動は、タンパク質の機能に関連するモーションと比較して、その大きさは小さい。すなわち、熱的な揺らぎによる座標変動のみが異なる 2 つの構造はその差異が小さく類似した構造をもつと考えられる。このことから、提案手法ではトラジェクトリ内の構造をクラスタリングし、熱的な揺らぎに由来する座標変動を除去することとした。

構造間クラスタリングによって、熱的揺らぎと比較的小さなモーションによる違いをもつ複数の構造を 1 つのクラスタとしてまとめることができる。

3.1 タンパク質構造間距離

タンパク質構造のクラスタリングを行うためには、構造間の距離とクラスタリングアルゴリズムが必要となる。ここではまずタンパク質構造間の距離について説明する。タンパ

Vol.2013-MPS-96 No.7
Vol.2013-BIO-36 No.7
2013/12/11

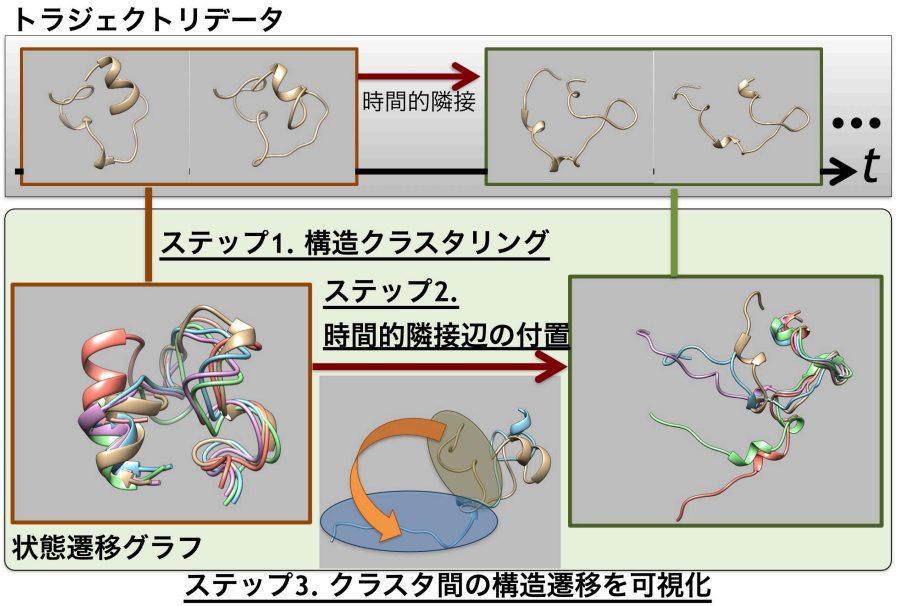

図 2 提案手法の概要

ク質構造間の距離を測るための指標として、各原子間のユークリッド距離に相当する Root Means Square Distance(RMSD) やタンパク質の 2 次構造の類似性を考慮した Dihedral Angle Distance(DAD) などが用いられてきた。しかし、RMSD はその計算を行うために比較対象となる 2 つのタンパク質構造間の回転および並進に対応する自由度を除去する必要があり、さらに三角不等式を満たさないためクラスタリング結果の評価が困難となる場合があること、また DAD においては DAD の変化とタンパク質構造の変化との間の相関が必ずしも高くないといった問題が指摘されている[13]。そこで本研究では、Distance Matrix Error(DME) という距離尺度を用いることとした。DAD では同一タンパク質内の各原子間のユークリッド距離の違いを考慮したものであるため、計算のための構造マッチングによる回転と並進の自由度を除去する必要がない、数学的な距離の公理である三角不等式を満たす、DAD と比較して構造的変化との相関が高い場合が多いといった利点がある。式 (1) にタンパク質構造 a と b の DME 距離 $DME(a,b)$ の定義を示す。

$$DME(a,b) = \frac{2}{N(N-1)}\sqrt{\sum_{i}^{N}\sum_{j\neq i}^{N}(d_{ij}^{a} - d_{ij}^{b})^2} \quad (1)$$

上記の DME を用いてトラジェクトリ内に含まれる構造のクラスタリングを行う。

3.2 Enhanced OPTICS アルゴリズムの概要

ある距離が与えられたとき、それを用いたクラスタリングには複数のアルゴリズムが存在するため、着目する問題に対して適切なクラスタリングアルゴリズムを選択する必要がある。ここでは本研究で使用したクラスタリングアルゴリズム選択の根拠について説明する。クラスタリングの方法としては、非階層的手法の代表的方法として k-means 法[14] や k-medioid 法[15]、階層的手法として Ward 法[16] がある。しかし、非階層的手法ではクラスタリングに入力として結果として出力されるクラスタ数を指定する必要があり、そのような事前知識がない場合には適用が難しいという問題点がある。この問題を解決するために、事前にクラスタリング数を指定する必要の無い x-means[17] という手法も提案されているが、この手法も点分布の形状を楕円形、すなわち Gauss 的分布と仮定しておりクラスタ形状が仮定された形状と異なる場合には、適切なクラスタリング結果が得られないという問題がある。同様に階層的クラスタリングである Ward 法においても、点分布によっては鎖状のクラスタが形成され、結果の評価が難しくなるという問題点がある。タンパク質がとりうる構造は、タンパク質構造を表現する空間 (Conformation 空間) におけるエネルギー地形の影響を受ける。すなわち、Conformation 空間上でエネルギー的に安定な領域に含まれる構造はトラジェクトリ内に多く含まれるが、構造的な過渡的状態に対応するようなエネルギー的に不安定な部分では、トラジェクトリに含まれる構造が少ないと考えられる。これは言い換えれば、トラジェクトリに含まれる構造群を Conformation 空間上にマッピングした際、点分布の密度が高い部分と低い部分が混在している可能性を示唆している。以上の考察から本研究では構造クラスタリングの手法として、密度を考慮したクラスタリング手法の 1 つである Enhanced OPTICS[18] を用いることとした。Enhanced OPTICS は従来のクラスタリング手法とは異なり、データのクラスタリングを直接行うのではなく、各データに対してデータ間の距離に基づいた順序を付置することによってクラスタリングを行う。具体的なステップは以下の通りである。まず第 1 に任意のデータを着目点として選択し、順序の 1 番目としてセットする。次に着目点から ϵ 以内のデータ集合 Neighbors を抽出する。このとき、Neighbors の要素数がある閾値 M 以下ならば着目点はノイズとして除去され、順序は付置されない。順序の付置を行った後、近傍集合 Neighbors と着目頂点

Vol.2013-MPS-96 No.7
Vol.2013-BIO-36 No.7
2013/12/11

```
1: procedure EnhancedOPTICS(データ集合 D, 近傍半径 ε, 有効近傍数 M)
2:     ClusterOrder ← φ
3:     UnprocessedPoints ← D
4:     for Point ∈ UnprocessedPoints do
5:         ClusterOrder ⇐ ExpandCluster(Point, ε, M)
6:     end for
7: end procedure
```

図 3 Enhanced OPTICS アルゴリズム

からの距離 Distance をリスト SortedList に格納する。ShortedList 内の要素が空でないなら、ShortedList から現在の着目頂点に付置された Distance 値に近い要素を抽出し、次の着目頂点とする。抽出された新たな着目頂点に順序を付置し処理済み頂点のラベルをつける。この SortedList からのデータ選択と近傍の Distance 値の計算、更新をすべてのデータが処理済みになるまで繰り返すことにより、与えられたデータ全てに対して順序を付置する。図 3 と図 4 に Enhanced OPTICS クラスタリングのアルゴリズムを示す。 図 3 と図 4 において、← は変数への代入操作、⇐ はリストへの挿入操作を示している。また、関数 **GetNeighbors** と **GetNextObject** はそれぞれ、引数で与えられた ObjectID から ϵ 以内に存在する要素 Neighbors とそれらとの距離 DistanceSet を返す関数と lastDistance に最も近い Distance 値を持つ要素を SortedList から選択する関数である。

上記アルゴリズムを用いて、トラジェクトリに含まれる構造をクラスタリングすることで、Conformation 空間のエネルギー地形に対応したクラスタリングが実現できると考えられる。

3.3 状態遷移グラフの構成法とモーション抽出法

Enhanced OPTICS アルゴリズムを用いた構造クラスタリングにより、各クラスタには互いに類似した構造がまとめらるため、クラスタ内の構造変動は PCA などの線形手法によって十分に近似できると考えられる。次に、形成されたクラスタ間を連結することにより、構造遷移を表現するグラフを構成する。具体的には、各クラスタをトラジェクトリに含まれる時間的隣接関係を用いて連結する。すなわち、各クラスタに含まれるある構造同士がトラジェクトリにおいて時間的に近いとき、対応するクラスタ間に時間的方向を考慮した有向辺を付置する。このように構成された有向グラフの各辺は、トラジェクトリ内で生じている大きな構造変化に対応していると考えられる。そこで、各クラスタを頂点、その間の時間

```
1: procedure ExpandCluster(データ ID ObjectID, 近傍半径 ε, 有効近傍数 M)
2:     OrderList ← φ
3:     ShortedList ← (ObjectID, NULL)
4:     Neighbors, DistanceSet ← GetNeighbors(ObjectID, ε)
5:     if |Neighbors| < M then
6:         return
7:     end if
8:     lastDistance ← NULL
9:     while SortedList is not Empty do
10:        ObjectID, Distance ← GetNextObject(lastDistance)
11:        lastDistance gets Distance
12:        OrderList ⇐ (ObjectID, Distance)
13:        Neighbors, Distance ← GetNighbors(SortedList, ObjectID, ε)
14:        SortedList ⇐ (Neighbors, DistanceSet)
15:    end while
16:    return OrderList
17: end procedure
```

図 4 関数 **ExpandCluster** の処理内容

的隣接関係を有向辺としたグラフをトラジェクトリに含まれる構造変化を記述しているグラフと捉え、これを構造遷移グラフと定義した。トラジェクトリの構造遷移グラフによる表現では、各クラスタ内におけるモーションは PCA などの線形手法で近似でき、クラスタをまたがる大きい、あるいは不連続的な変化は有向辺として表現されることから、急激な構造変化を含むトラジェクトリや不連続的なサンプリングから得られたトラジェクトリでもモーションを十分に表現できる。

状態遷移グラフの各辺において生じている構造変化を抽出することにより、トラジェクトリに含まれる状態遷移を経時的に抽出することができる。本研究では、構造遷移グラフの各辺に対応する構造変化を UCSF Chimera[19] を用いて可視化することによって抽出することとした。

このようなクラスタリング手法とグラフを組み合わせたタンパク質モーションの記述は本研究より以前にもいくつかの報告において行われている[20][21]。しかし本提案とそれらの報告

Vol.2013-MPS-96 No.7
Vol.2013-BIO-36 No.7
2013/12/11

では、Rajan[20] による手法が従来の階層的クラスタリングを用いているため、Conformation 空間内のエネルギー地形を反映していない可能性があること、クラスタ間の辺を付置する際に Conformation 空間の近接性のみに着目しているため、実際にはトラジェクトリに含まれていない辺がグラフに含まれている可能性があり、構成されたグラフがタンパク質の真のモーションを表現していない場合がある点が異なる。また、Chiang[21] による手法は予測を主な目的にしており、Markov 条件を満たす結果を得るために用いるトラジェクトリに対して十分なサンプリング条件を科しているという点が本提案手法と異なる。

4. 適用実験

本提案手法の有効性を示すために、Villin headpiece subdomain[22] (HP-35 NleNle) のトラジェクトリに対して提案法の適用実験を行った。

実験に用いたトラジェクトリは Villin headpiece subdomain(HP-35 NleNle) というニワトリのタンパク質のシミュレーションデータである。HP-35 NleNle は、それ単独で折り畳みが進むこと、さらに折り畳みが他のタンパク質と比較して早く生じることから、折り畳みのモデルとして良く研究されている。HP-35 NleNle のトラジェクトリは報告[22] で構築されたものを用いた。以下に使用したトラジェクトリの生成方法の概要を示す。トラジェクトリ生成法の具体的なパラメータについては報告[22] を参照されたい。トラジェクトリの生成においては、Villin headpiece の X 線結晶構造 (PDB ID 2F4K) から AMBER2003 力場[23] のもと、温度を 373K に設定して事前シミュレーションを行い合計 9 つの折り畳まれていない初期構造が生成された。これらの初期構造に対して、温度を 300K に設定してシミュレーションが実行された。シミュレーションにおける初期速度は Maxwell-Boltzmann 分布からランダムに各原子に付置された。各シミュレーションでは、50ps 毎の座標計測が 400step 行われた。本研究では、9 個の異なる初期構造から生成された 50ps × 400step = 20ns の長さの 9 トラジェクトリを 400ps 毎にサブサンプリングを行い、実験に使用するデータを生成した。

図 5 に 9 つのトラジェクトリから生成された構造遷移グラフを示す。図 5 において、RunX が初期構造 X から得られたトラジェクトリに対応する結果を示している。次に非線形なタンパク質モーションが抽出可能であるかを確認するために、構造遷移グラフの有向辺に対応する遷移の可視化を行った。可視化の例として、Run0 の四角で示された部分の辺の可視化結果を図 6 に示す。 図 6 から明らかなように、タンパク質の中央付近の部分的回転や端付近の回転といった、非線形なモーションが抽出されていることが確認できる。

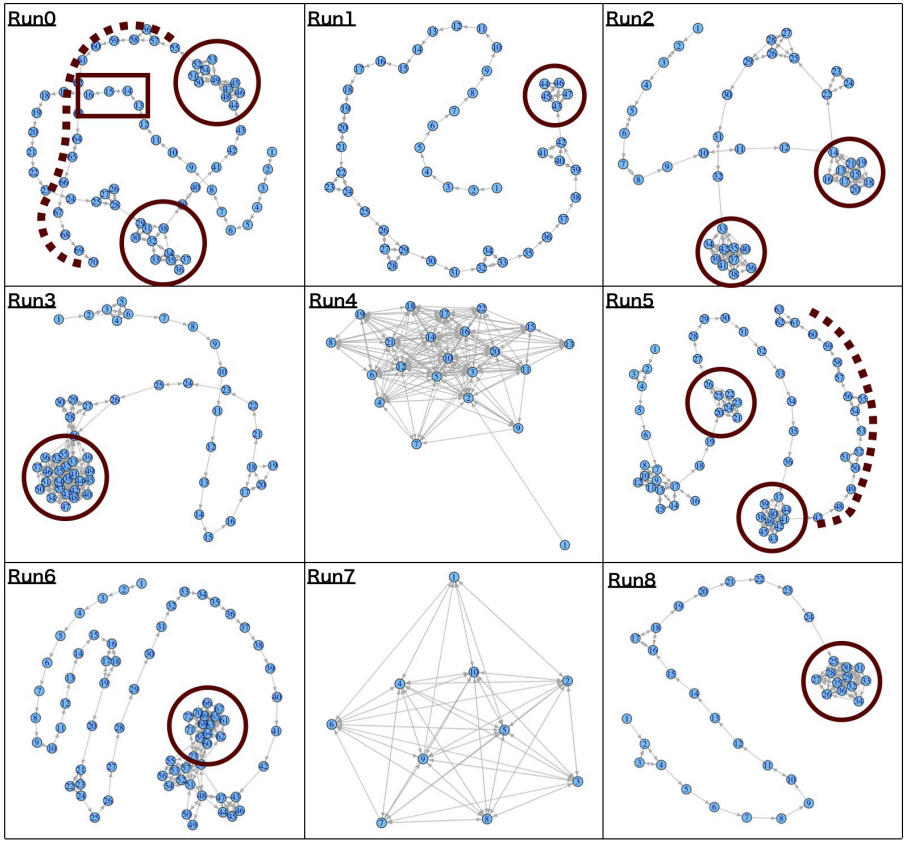

図 5 9 つのトラジェクトリから得られた構造遷移グラフ

Vol.2013-MPS-96 No.7
Vol.2013-BIO-36 No.7
2013/12/11

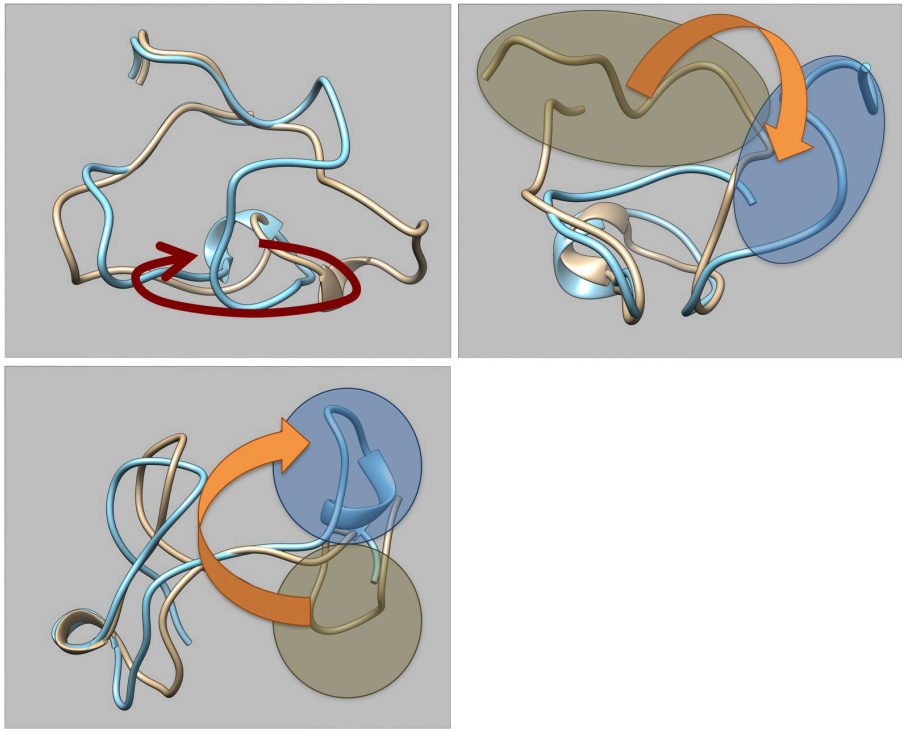

図 6 Run0 の状態遷移辺の可視化例: ベージュで示された構造が遷移前、シアンで示された構造が遷移後の構造を表す。

次に従来の報告との整合性を調べるために、構造遷移グラフの構造を調べた。結果から、Run4 と Run7 における状態遷移グラフの頂点数が他のグラフと比べて少ないことが分かる。各クラスタが類似した構造をまとめたものであることから、トラジェクトリに含まれる構造が少ないと考えられる。この結果は Ensign らの報告[22] において、初期状態 4 と 7 から得られたトラジェクトリが他と比べて折り畳みが速やかに進んだという報告と対応している。さらに、Run4 と Run7 の状態遷移グラフを比較すると Run7 のほうが辺密度が低いことが分かる。これは、クラスタ間の状態遷移が少ないという点で Run4 よりも Run7 のほうがより速く折り畳みが進行したという同報告の結果と合致している。一方で、これらの状態遷移グラフの中には赤い円で示されるような辺が密集している領域があることも確認される。これは構造遷移がある時間においてトラップされた状態に対応している。Run1 は Ensign らの報告[22] によれば、折り畳み状態に到達しなかった唯一の結果だが、状態遷移グラフをみるとその状態遷移が最後にはトラップされた状態にあることが分かる。また、Run0 と Run5 では辺の密集領域から破線で示されるような状態遷移がさらに進んでいることが確認された。これらの経路では、始端となる円で示された部分が折り畳み状態になっており、その折り畳み状態からの折り畳まれていない状態へと構造が遷移していることが分かった。

5. 結　　論

本研究では、タンパク質モーションのシミュレーション結果から、従来手法では扱いが難しかったタンパク質の非線形モーションの抽出手法の提案を行った。提案法では、トラジェクトリに含まれる構造が Conformation 空間のエネルギー地形に対応していることを考慮して、密度ベースクラスタリング手法の 1 つである Enhanced OPTICS アルゴリズムを用いてクラスタリングを行い、クラスタ間の遷移をトラジェクリ内の時間的隣接関係から有向辺として与えることにより、トラジェクトリに含まれる構造変化をグラフとして表現した。提案法の有効性を確認するために、HP-35 NleNle のトラジェクトリデータに対して提案法の適用を行った。その結果、タンパク質の非線形なモーションがタンパク質モーションを表現する状態遷移グラフにおける有向辺として抽出可能であることが確認された。また、抽出された状態遷移グラフの構造が HP-35 NleNle の折り畳みに関する既存研究の報告内容[22] と合致することが確認された。さらに、既存報告では報告されていなかった、折り畳みに至らないトラジェクトリに対して、構造遷移におけるトラップ状態が存在していることが確認された。

Vol.2013-MPS-96 No.7
Vol.2013-BIO-36 No.7
2013/12/11

今後の課題としては、現在マニュアルで行っている構造遷移の抽出の自動化が挙げられる。

参 考 文 献

1) Demirel and Keskin, “Protein Interactions and Fluctuations in a Proteomics Network using an Elastic Network Model”, *J. Biomol. Struct*, **272**, pp.381-386, 2005
2) Tssell, R.T. and Callis, P.R., “Simulations of Tryptophan Fluorescene Dynamics during Folding of the Villin Headpiece”, *J. Phys. Chem.*, B, **116**, pp.2285-2294, 2012
3) Zhang, J.L, Zheng, Q.C, Li, Z.Q and Zhang, H.X, “Molecular Dynamics Simulaitons suggests Ligand’s Binding to Nicotinamidase/Pyrazinamidase”, PLos ONE, **7**(6), e39546, 2012
4) Karr, J.R, Sanghvi, J.C., Macklin, D.N., Gutschow, M.V, Jacobs, J.M., Bolival, B., Assad-Garcia, Glass, J.I. and Covert, M.W., “A Whole-Cell Computational Model Predicts Phenotype from Genotype”, *Cell*, **150**(2), pp.389-401, 2012
5) Serpell, L.C., “Alzheimer’s Anlyloid Fibrils: Structure and Assembly”, *Biochem Biophys Acta*, **1502**, pp.16-30, 200
6) Alder, B.L, and Wainwright, T.E, “Studies in Molecular Dynamics I. General Method”, *J. Chem. Phys.*, **31**(2), 1959
7) D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Goetz, I. Kolossvry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman, AMBER 13, University of California, San Francisco, 2012
8) Pronk, S., Pà ll, S., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M.R., Smith, J.C., Kasson, P.M., Spoel, D., Hess, B. and Lindahl. E, “GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit”, *Bioinformatics*, **29**(7), pp.845-854, 2013
9) Sugita, Y. and Okamoto, Y., “Replica-Exchange Molecualr Dynamics Method for Protein Folding”, *Chem. Phys. Letter*, **314**, pp.141.151, 199
10) Pande Lab, “folding.stanford.edu”, Stanford University, 2012
11) Amadei, A., Linssen, A.B.M. and Berendsen, H.J.C, “Essential Dynamics of Proteins”, *Proteins: Struct. Funct. Genet*, **17**, pp.412-425, 1993
12) Tenenbaum, J.B., Silva, V. and Langford, J.C, “A Global Geometric Framework for Nonlinear Dimansionality Reduction”, *Science*, **290**, pp.2319-2323, 2000
13) Fuglebakk, E., Echave, J. and Reuter, N., “Measuring and Comparing Structural Fluctuation Patterns in Large Protein Datasets”, *Bioinformatics*, **28**(19), pp.2431-2440, 2012
14) , Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y., “An Efficient *k*-Means Clustering Algorithm: Analysis and Implementation”, IEEE Transactions on Pattern analysis and Machine Inteligence, **24**(7), 2002
15) Liu, X. and Sheng, W., “A Hybrid Algorithm for *k*-Medioid Clustering of Large Data Sets”, IEEE Evolutionary Computation, **2**, 2004
16) Bieniasz, A. and Majchrzak, A., “Applying the Ward Method in the Analysis of Financial Simulation of Commercial Banks”, *e-Finanse: Financial Internet Quarterly*, **7**(3), pp.1-12, 2011
17) Pelleg, D. and Moore, A., “X-Means: Extending K-means with Efficient Estimation of the Number of Clusters”, *Proceeding of the 17th International Conf. on Machine Learning*, pp.727-734, 2000
18) Bendre, M., “EOPTICSClus: Enhanced OPTICS based Clustering”, CS512 Course Final Presentation, 2012
19) Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., ferring, T.E., “UCSF Chimera: A Visualization System for Exploratory Research and Analysis”, *J. COmput. Chem*, **25**(13), pp.1605-1612, 2004
20) Rajan, A., Freddolino, P.L. and Schulten, K., “Going beyond Clustering in MD Trajectory Analysis: An Application to Villin Headpiece Folding”, PLoS ONE, **5**(4), e9890, 2010
21) Chiang, T.H., Hsu, D. and Latombe, J.C., “Markov Dynamics Models for Long-Timescale Protein Motion”, *Bioinformatics*, **26**, pp.i269-i277, 2010
22) Ensign, D.L, Kasson, P.M. and Pande, V.S., “Heterogeneity even at the Speed Limit of Folding: Large-Scale Molecular Dynamics Study of a Fast-Folding Variant of the Villin Headpiece”, *J. Mol. Bio.*, **374**, pp.806-816, 2007
23) Wang, J.M., Cieplak, P. and Kollman, P.A., “How Well Does a Restrained Electrostatic Potential(RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules?”, *J. Comput. Chem.*, **21**, pp.1049-1074, 2000