社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (59)
2007－BIO－11 (59)
2007／12／21

確率的情報処理による生体分子の熱揺らぎ解析に関する研究

徳永慎一[†] 関嶋政和[††,†††]
村岡洋一[†††] 野口保[††,†††]

近年、計算機の発展によりスーパーコンピューターやＰＣクラスタを使うことによって、細かいタイムステップでもナノ秒単位の生体分子のシミュレーションが短時間で出来るようになった。それに伴い、大量のデータの解析には様々な手法が用いられている。一般的なシミュレーションデータ解析では、タンパク質の状態遷移を可視化することが多く、それらはタンパク質がある構造から別の構造に遷移する現象を明らかにすることを目指している。本研究では分子動力学シミュレーションで得られたトラジェクトリデータをいくつかの構造に分類し、HMM(隠れマルコフモデル) による解析を行うことで、タイムステップ毎の時系列遷移からは解析が難しい、タンパク質の熱揺らぎのパターン解析を提案する。また、マイクロ秒に及ぶシミュレーションにより得られた膨大なデータを円滑に解析することができる一貫したシステムの開発も行った。

Analysis of protein folding using probabilistic information processing

Shinichi Tokunaga,[†] Masakazu Sekijima,[††,†††]
Yoichi Muraoka[†††] and Tamotsu Noguchi[††,†††]

Rapidly increasing computational power enables to relatively long (10 to 100 nano seconds) molecular dynamics simulation in a realistic time. As a result of huge trajectory data through simulation, varieties of methods are applied to them. General analytical approaches attempt to visualize transition states of trajectory to show local minima of energy landscape, because they are important to understand the thermodynamics and kinetics of protein folding. In this work, we clustered trajectory data of molecular dynamics simulation and applied HMM (Hidden Markov Model) to reveal transition state pattern of protein folding. In addition, we developed a system that could analyze huge amounts of simulation data.

1. はじめに

タンパク質は生体分子の中で最も重要な物質の一つであり、その機能の発現には固有の立体構造の形成が必要である。化学的には鎖状の高分子であるタンパク質が、どのような状態を経由して折りたたみ、最終的にどのような立体構造をとるのかという疑問はフォールディング問題と呼ばれる。天然状態においてはタンパク質は非常に安定した状態に折りたたまれて(フォールディング)いることが多く、なんらかの機能をもっていることが多い。つまり、そのタンパク質の安定な状態や存在確率の高い状態を突き止めることは、タンパク質の構造・機能予測につながる。これを読み解くために分子動力学シミュレーションが用いられている。

分子動力学法は、生命科学の分野において生体分子のダイナミクス、機能や構造予測を解析するために用いられている。分子動力学法は主に溶液中の生体分子をシミュレートするのに用いられるが、一般に生体分子のシミュレーションは、大規模で複雑な系を扱い、しかも生体内に近い環境での精密な計算が必要であることから、多大な計算時間を必要とする。こうした分子動力学法のような膨大な計算時間を必要とするシミュレーションに対して、並列計算は必要不可欠な技術となっている。

近年、計算機の発展によりスパコンや並列クラスタを使うことによって、ナノ秒単位のシミュレーションが短時間で出来るようになった。しかし、100ナノ秒～マイクロ秒単位のシミュレーションを行うにはまだまだ時間がかかり、マイクロ秒単位のシミュレーショ

† 早稲田大学大学院理工学研究科 情報ネットワーク専攻
Graduate School of Science and Engineering, Waseda University
†† 産業技術総合研究所 生命情報工学研究センター
Computational Biology Reserch Center, Natinal Institute of Advanced Industrial Science and Technology
††† 早稲田大学理工学術院
Faculty of Science and Engineering, Waseda University

ンによる大量データの解析例は少ない。

分子動力学シミュレーションで得られたデータの解析には、各トラジェクトリデータの Cα(またはバックボーン) 間の二面角を用いることが多いが、本研究では Cα 間の距離を用いて解析を行う。

また、一般的なシミュレーションデータ解析では、タンパク質の状態遷移を可視化することが多く、それらはタンパク質がある構造から別の構造に遷移する現象を明らかにすることを目指している。タンパク質の熱揺らぎのパターン解析には自由エネルギー地形や統計処理などを用いた解析手法が存在するが、情報工学の観点から解析を行った例は少ない。したがって、分子動力学シミュレーションで得られたトラジェクトリデータをいくつかの構造に分類し、確率情報処理の観点からトラジェクトリデータの状態遷移を可視化することは、タンパク質の安定な状態や存在確率の高い状態を明らかにする一つの指針として、有用な手段であると考えられる。そこで、本研究ではトラジェクトリデータを HMM(Hidden Marcov Model) にかけることで、シミュレーションにおけるタンパク質の状態遷移を確率的なモデルとして可視化することにより、タイムステップ毎の時系列遷移からは解析が難しい、タンパク質の熱揺らぎのパターン解析を提案する。HMM は確率モデルの一つであり、「システムがパラメータ未知のマルコフ過程である」と仮定し、観測可能な情報からその未知のパラメータを推定することができる。音声認識、ゲノミクス、形態素解析 (自然言語処理) などに応用されており、不確定な時系列のデータをモデル化するための有効な確率統計的手法である。

また、マイクロ秒単位のシミュレーションデータの解析を行うにあたり、一般の PC ではメモリ容量がボトルネックとなる。たとえば、倍精度浮動小数点 (64bit=8byte) で表現された 1,000 × 1,000 × 1,000 の配列は約 8GB となり、32bit マシンが持つメモリ空間 4GB を越えてしまうため、大きなデータを扱うには限界がある。そうなると、データはファイルとして記録するか、もしくはデータベースに記録しておき、その一部をメモリへと読み込んで処理しなくてはならなくなる。本研究ではシミュレーションで得た膨大なデータを扱うにあたり、1PC でも解析が円滑に行えるシステムを提案する。

したがって本研究では、1.) タンパク質シミュレーションデータの解析、2.) 分子動力学シミュレーションのデータ解析を行う一貫したシステムの構築の 2 つを目的とする。

本稿の構成は以下の通りである。2 章では本研究で行った分子動力学シミュレーションについて説明する。次に, 3 章では解析を行うシステムについて説明する。4 章では、解析結果を述べ、5 章で全体をまとめる.

表 1 データセット

シニョリンタンパク質 (PDBID:1UAO)
10 residues (138 atoms)


2. 分子動力学シミュレーション

分子動力学法とは、生体分子のような多分子系において、ニュートンの運動方程式を積分することによって、個々の原子の運動をシミュレーションする手法である。このような積分は多数の原子が互いに相互作用しながら運動しているので多体問題となり、解析的に解くことが事実上不可能であるため、有限差分法を用いて数値的に解くことになる。

分子動力学法の基本手順は粒子の位置や速度を短い時間刻み Δ t で離散的に計算していくというものである。すなわち、ある時刻において原子間に働く力を計算し、それに基づいて Δ t 後の原子の位置を求め、その求めた位置で新たに原子間に働く力を計算する、という操作の繰返しである。温度や圧力を一定にする場合、その調整のための操作も繰返しに含まれる。こうして得られた原子の座標の軌跡をもとに、種々の物理量を計算する。

2.1 シミュレーション設定

本研究では、分子動力学シミュレーションプログラム SANDER (Simulated Annealing with NMR-Derived Energy Restraints)[5] を産業技術総合研究所生命情報工学研究センターの IBM Blue Gene/L 上で実行した。解析例としたシニョリンタンパク質[6] は、産業技術総合研究所の本田真也博士らによって設計・合成された世界最小のタンパク質 (表 1) である。ポテンシャル関数には AMBER の ff99[7] を用い、二面角には Simmerling[8] の補正を行っている。水の作用には、Generalized Born モデル[9]、温度は 300K、MD の 1step を 1 フェムト秒とし、NPT アンサンブルで

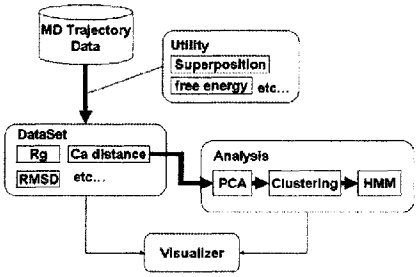

図 1 システム構成

1 マイクロ秒のシミュレーションを Blue Gene/L の 32 ノード（64CPU）を用いて 1433 時間行った。

3. 解析システム

MD により得られたトラジェクトリデータを、自由エネルギーの地形や、生体分子の周りの水分子の動き、PDB(Protein Data Bank) 構造との RMSD など様々なデータとして可視化することができるシステムを考案する。図 1 に示すように全体のシステムは、DataSet、Analysis、Visualizer の大きく 3 つのモジュールに分けることができる。

3.1 DataSet

主に入力データと出力データを保持するモジュールである。MD で得たトラジェクトリデータは、座標データの他にも以下のような様々なデータとして保持される。

- Rg - コンパクトさ
- RMSD - 入力構造との RMSD
- Cα distance - 各構造の Cα 間距離の全組み合わせ

Utility モジュールにはタンパク質同士の重ね合わせ (Superposition) に関するプログラムや、自由エネルギー地形を描くためのプログラムが含まれている。

3.2 Analysis

解析を実際に行うモジュールである。現在の実装では Cα distance を入力として解析を行う。

- PCA
  各構造について主成分分析を行い、第 1〜3 主成分までを抽出
- Clustering
  k-means 法による非階層クラスタリングを行う
- HMM
  各クラスターを一つの状態とみなし、HMM にかける。HMM の学習には Baum-Welch アルゴリズムを用いる。

Cα 間距離を主成分分析することによって、シミュレーション中に存在確率の高かった構造を抽出することができる。また、第 1〜第 3 主成分を座標軸とした PCA 空間をクラスタリングによってグループ分けすることにより各クラスターを一つの状態とみなし、シミュレーション内におけるタンパク質をいくつかの構造に分類する。さらに、各クラスター (構造) を入力として HMM にかける。

3.3 Visualizer

各解析結果を表示するモジュールである。

- 第 1〜第 3 主成分の PCA 空間の投射
- 各保持データを座標軸としたヒストグラム
- 時系列に沿った状態遷移図
- 確率的状態遷移図

クラスタリングの結果より、トラジェクトリデータの状態が時系列に沿ってどの状態により多く存在しているのか、どの状態からどの状態により多く遷移しているのかを可視化することができる。また、HMM にかけた結果により確率的な状態遷移を可視化できる。

3.4 大規模データ処理

本研究で扱うシミュレーションデータは 1 マイクロ秒という膨大な量のデータである。構造サンプリングに用いたタイムステップは 20 フェムト秒なので、入力ファイルである時系列データは 50000 ファイルにも及ぶ。全データについての主成分分析を行うと、共分散行列を作成するのに少なくとも 50,000 × 50,000 の行列演算処理が必要である。この行列を倍精度浮動小数点の配列で表現したとすると、少なくとも約 25GB のメモリが必要となりシステムにおけるボトルネックとなってしまうため、1PC で解析を行うには現実的ではない。よって大規模データを扱う上での本システムの方針を以下に示す。

1 時系列データの中から飛び飛びのデータだけを抽出する。例えば 20 フェムト秒ではなく 40〜100 フェムト秒ごとのデータだけについて解析を行う。

2 Clustering までの処理を行い、存在確率の高い構造を抽出する。

3 各クラスターに存在する構造の平均構造を取る。

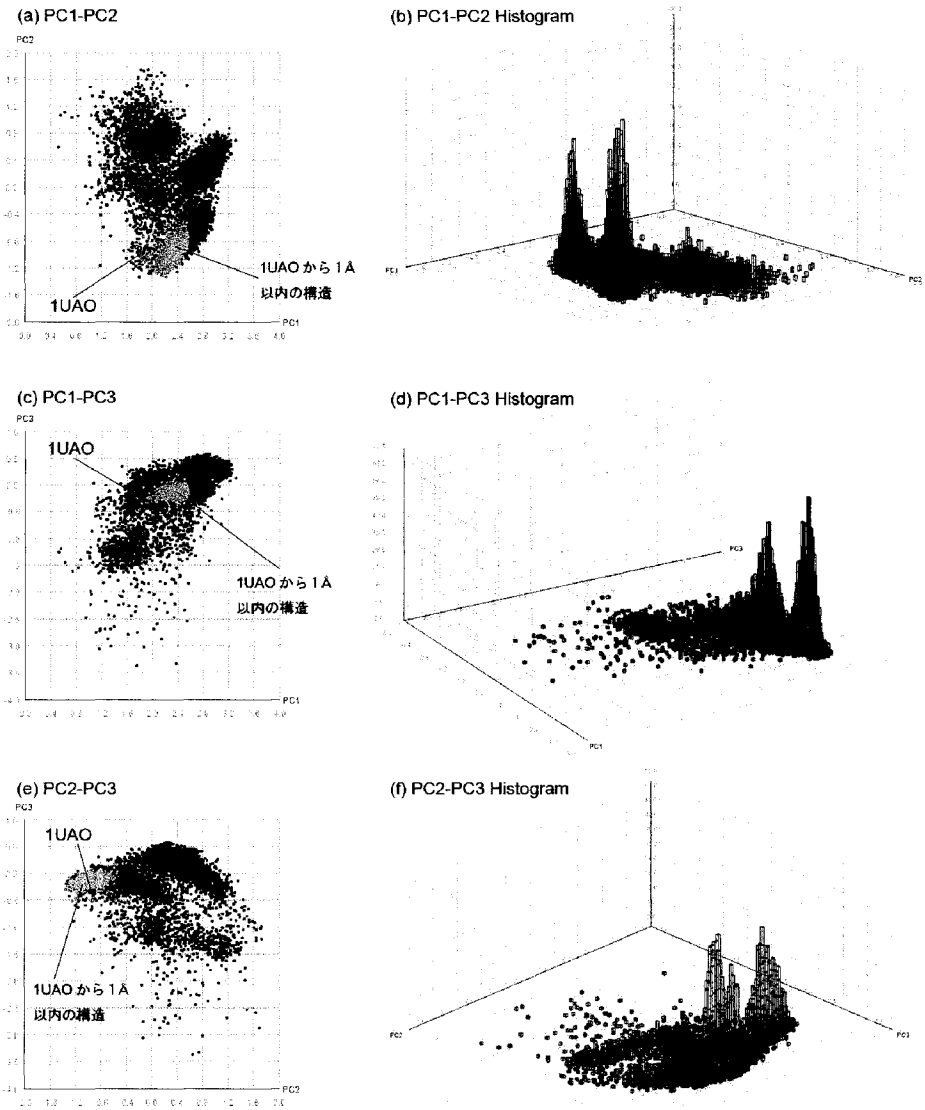

図 2 Cα distance を PCA にかけた結果

4 3 で抽出した平均構造との RMSD の低い値だけを持つ入力データを抽出し、再び解析を行う。

1 の処理は、入力データがメモリ容量に乗るデータ量になるように動的に行われる。2 の処理でのクラスターの個数、3 の処理での RMSD の値については手動で閾値を設定する。

4. 解析結果

2 章で行ったシニョリンタンパク質のシミュレーションデータを、本システムにより解析した結果を示す。図 2 は、各トラジェクトリデータの Cα 間距離を主成分分析し、第 1、第 2 主成分抽出したものである。図 (a)、(c)、(e) については、シニョリンの PDB 構造である 1UAO と、1UAO から RMSD が 1.0Å 以下の構

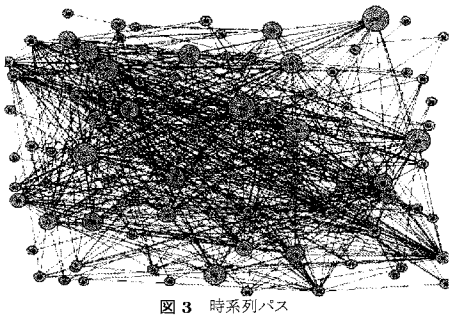

図 3 時系列パス

造についても示してある。それぞれのヒストグラムと比較してみると、シミュレーション中もっとも存在確率の高いタンパク質構造ではなく 2 番目に存在確率の高い場所に、PDB 構造とそれに近い構造が存在しているという結果が得られた。

図 3 は第 3 主成分まで分析した結果を k-means によりクラスタリングした結果である。一つ一つのクラスターは円で表されており、円の大きさが各クラスターに含まれるトラジェクトリデータの個数を表している。円と円の間に結ばれている線は、各クラスター内のあるトラジェクトリから、他のクラスター内のトラジェクトリへの時系列遷移を表している。図に示してあるクラスターの数は全部で 100 個であり、黒く塗りつぶしてあるものが PDB 構造のあるクラスターである。

5. ま と め

本稿では、生体分子の熱揺らぎ解析を行うシステムの開発を行い、また、本システムを用いて主成分分析からクラスタリングまでの解析を行った結果について述べた。このシステムでは MD シミューレヨンにより得られたトラジェクトリデータを、DataSet モジュールによって Rg、自由エネルギー、RMSD などの解析しやすいデータとして保持しておく。Analysis モジュールによって主成分分析、クラスタリング、HMM による解析を行い、Visualizer モジュールによって DataSet、Analysis で得たデータを可視化することができる。さらにこのシステムは、PC のメモリ容量だけでは処理しきれない大規模なデータを扱う場合であっても、クラスタリングによるデータのフィルタリングを行うことで、高速な処理を行える解析機構を備えている。

今後の課題として以下の事柄が挙げられる。

1.) クラスタリング結果の HMM による解析
クラスタリングによって分類された各構造を HMM にかけることにより、タンパク質シミュレーション内の確率的な構造遷移をモデル化する。

2.) 他のタンパク質シミュレーションへの適用
シニョリンのシミュレーションデータを解析したが、シニョリンはわずか 10 残基の世界最小のタンパク質であり、シミュレーション中の各構造同士の RMSD の値はどんなに大きくても 5.0Å を超えることはない。したがって本システムの精度をあげるために、生体分子がより大きく揺らぐようなタンパク質の解析例が必要だと考えられる。

3.) 様々なデータを入力とした解析モジュールの開発
本稿では Cα distance を用いた解析システムの開発を行ったが、Cα 間の二面角などを用いた解析も選択的に行えるようにしたい。

4.) さらなる大規模データ処理の検討
主成分分析からクラスタリングまでかけたデータを RMSD の閾値でフィルタリングする手法について述べた。しかし、より正確な解析を行うためには全データを同時に解析できる必要がある。現在の提案システムよりはパフォーマンスが下がる可能性があるが、RAID0(HDD ストライピング) や Flash Memory を用いたディスクアクセスの高速化を行ったのち、データベースやファイルをメモリ代わりに使用するシステムの作成を考えている。

参 考 文 献

1) Sergei V. Krivov and Martin Karplus : Hidden complexity of free energy surfaces for peptide (protein) folding, PNAS, Vol.101, No.41, pp. 14766–14770 (2004).
2) David A. Evans and David J. Walesa : Folding of the GB1 hairpin peptide from discrete path sampling, J.Chem.Phys., Vol.121, No.2, pp. 1080–1090 (2004).
3) Kelly L. Damm and Heather A. Carlson: Gaussian-Weighted RMSD Superposition of Proteins: A Structural Comparison for Flexible Proteins and Predicted Protein Structures, Biophys.J., Vol.90, No.12, pp.4558–4573 (2006).
4) Kauyoshi Ikeda, Kentaro Tomii, Tsuyoshi Yokomizo, Daisuke Mitomo, Keiichiro Maruyama, Shinya Suzuki and Junichi Higo : Visualization of conformational distribution of short to medium size segments in globular proteins and identification of local structural motifs, Protein Sci, Vol.14, No.5, pp.1253–1265 (2005).

5) D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, D.A. Pearlman, M. Crowley, R.C. Walker, B. Wang, S. Hayik, A. Roitberg, G. Seabra, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D.H. Mathews, C. Schafmeister, W.S. Ross, and P.A. Kollman, 2006, AMBER 9, University of California, San Francisco.
6) S. Honda, K. Yamasaki, Y. Sawada, and H. Morii, 10-residue folded peptide designed by segment statistics, *Structure*, 12, 2004, pp. 1507-1518.
7) Wang, J., Cieplak, P. and Kollman, P. A. (2000) J. Comput. Chem. 21, pp.1049-1074
8) Simmerling, C., Strockbine, B. and Roitberg, A. E. (2002) J. Am. Chem. Soc. 124, pp.11258-11259
9) Tsui, V. and Case, D. A. (2001) Biopolymers 56, pp.275-291