10-MD02

生体分子の大規模シミュレーションにより得られる時系列データの解析

戸田幹人（奈良女子大学）

概要　　たんぱく質など生体分子の分子動力学計算で得られる大量の時系列データから，分子機能と関連する集団運動が成す階層的動力学の情報を抽出する．そのため，カーネル法を始めとする統計科学的な解析手法を発展させる基礎的な研究，および，基礎的研究で展開される手法を実際の大規模データの解析に適用するためのシステム開発を目的とした共同研究について報告する．九州大学のスーパーコンピュータシステムを利用した時系列データ取得のための大規模計算と，時系列解析手法としてのカーネル法を始めとする解析手法の確立，および，それらの手法を大規模データに応用できる用に実装するデータ処理システムの構築の状況について，これまでの成果と合わせて現状を述べる．

1.　研究の目的と意義

揺らぐ環境の中にある生体分子の制御という将来の目標を想定し，この目標への第一歩として動力学計算で得られる時系列データの解析によって，分子のダイナミクスを深く理解することを目的とする．同時に，このように複雑な物理系の時系列を得るための大規模数値計算手法の利用と，計算によって得られる時系列データから意味のある物理学的知見を抽出するための数値データの情報処理的手法の確立を目指す．

従来，階層的な系に対する「マルチスケールシミュレーション」では，それぞれの階層に対応するアルゴリズムをつなげることにより，大規模なシミュレーションを実現してきた．これらの手法の発展の中で，「マルチスケールシミュレーション」の基礎理論の必要性が高まっている．特に階層的集団運動の動力学的メカニズムにさかのぼった議論は従来ほとんどなく，計算科学・非線形科学・統計科学の共同によるブレークスルーの必要性が大きい．本研究では大自由度力学系の相空間構造，特に「ゆっくりした集団運動」の動力学を記述する「不変集合」に基づく理論に基いて，階層性の起源に立ち返る解析を行う．また，量子力学的な振動モード解析から構築される分子のモデルと，階層的集団運動の動力学に依拠した新たな「粗視化」モデルの比較・検討も将来的な目的としている．

本研究では生体分子のダイナミクスを理論的・計算的アプローチによって両面から調べるが，従来の研究は小さな分子を詳細に調べるか，大きな分子を古典力学的に調べるかのどちらかに偏っていたため，本研究はその間を埋める新しい領域での試みとなる．このため本研究は，必然的に物質科学系の物理・化学・生物などさまざまな分野への貢献が期待される．この研究の過程で，大規模な計算機を利用した数値シミュレーションにより動力学的データを取得し，この多次元時系列情報を含む大量の数値データから意味的に重要な情報を抽出するという処理が必要とされるため，ハイパフォーマンスコンピューティングやデータマイニング・時系列情報処理などの情報科学分野の最先端の方法論を利用・開拓することとなる．

特に，数値計算による時系列に対するデータマイニングに関して，これまでの研究経過と最近の発展から本研究の目的と意義を詳述しておく．

われわれは，タンパク質分子の分子動力学データに対し，ウエーブレットにより集団運動を抽出した予備的な計算を行ってきている(大林・戸田，分子動力学データは横浜市大の木寺・渕上から提供していただいた)．これにより，ウエーブレットを使ってゆっくりした振動成分を抽出すると，時々刻々に変動する集団運動的な振動運動が取り出せることがわかっている．このように抽出された集団運動の動的な挙動の解析に向けて，非線形主成分解析・独立成分解析・カーネル相関解析・

特異値分解・ランダム行列理論[Kubotani, Toda and Adachi, Phys. Rev. A**74**, 032314 (2006); Kubotani, Adachi and Toda, Phys. Rev. Lett. **100**, 240501 (2008), Adachi, Kubotani and Toda, Annals of Physics, **324**, 2278-2358 (2009)]などの手法を駆使し，集団運動の動力学を抽出する「データマイニング」の方法が確立できるものと考えている．そこで，共同研究者の福水健次(統数研)の研究成果であるカーネル法に依拠した統計科学的解析(カーネル法入門:正定値カーネルによるデータ解析，朝倉書店,2010)を応用する形で，複数の階層に渡る集団運動間の相互作用に対して，統計的因果推論を利用した解析を試みる．階層的集団運動の動力学は，閉じた微分方程式で記述される「決定論」ではなく，集団運動以外の自由度から影響を受ける「ランダム力学系」であり，時系列データから「ランダム力学系」を構築するには，揺らぐ系における「因果関係」の推論が必要である． 統計科学では従来，時系列の因果性の解析は，定常性など限定的な状況でなければ困難と思われてきた．しかし近年，福水らはカーネル非線形回帰分析に基く因果推論を展開，Max-Planck研究所(Tuebingen)や Peter Spirtes のグループなどもその適用を始めており[R. Tillman, A. Gretton, P. Spirtes, Proc. NIPS (2009)]，今後，有望な方法論になると考えられる．一般に，実験できない状況で得られるデータに対して，従来の統計科学では，相関関係の解析はできるが因果関係の推論はできないとされてきた．この点で重要な貢献をしたのが，Judea Pearl や Peter Spirtes, Clark Glymour らである．彼らは因果関係の厳密な定式化により，統計解析から因果関係の推論ができる場合があることを示した．他方で経済時系列では，定常AR過程に基づく Granger Causality がよく用いられてきた．本研究ではこれらを非線形非平衡な時系列データに拡張し，階層的集団運動の時系列データから階層間相互作用や外界の刺激との因果関係の解析を行う．非平衡な時系列データの統計的因果推論そのものが新しく，その分子動力学データへの応用は意義深い．

これまでも，物質系科学と情報系科学の境界領域での研究は数多く実施されてきているが，物質系科学に重点を置く場合には，情報系科学として開発された結果のみを利用することが多く，逆に情報系科学に重点を置く研究では，物質系科学を単に実世界での必要性をアピールするために利用することが多かった．本研究は，物質系科学における最先端の理論的研究に軸足を置きつつも，数値時系列データからの情報の抽出という点で，情報科学においても新たな手法の構築を必要とするものになっている．実際の大規模生体分子から得られる時系列に対して，大規模データのマイニングや統計的因果推論，不変集合に基づく階層的アルゴリズムの開発を行った例はまだないため，これが本研究の特色である．これらの学際的な共同研究の成果によって新しい計算・解析方法が確立し，その結果として生体機能の理解を深めることが可能となる．

2. 当拠点公募型共同研究として実施した意義

(1) 共同研究を実施した大学名
九州大学
(2) 共同研究分野
超大規模数値計算系応用分野，および，
大規模データ処理系応用分野
(3) 当公募型共同研究ならではという事項など
比較的安価にスーパーコンピュータが利用できたこと．シンポジウムなどで，他の分野の研究者と交流が出来たこと．

3. 研究成果の詳細

これまでの研究は，小規模の時系列データに対する統計的解析の適用とその有効性の検証を中心に実施してきている．この中間報告では，検証に利用したデータの規模と，その結果について，報告する．また，上記有効性の検証とは別に，時系列データ解析システムとして開発中のソフトウェアについても，現時点までの開発内容と，システ

ムの概要に関して報告する．

では，最初に時系列データ処理の検証について，報告する．時系列データ処理に関して，小規模のデータに対する検証は主に下記の２点で成されている．一つはタンパク質の分子動力学データにおいて，比較的短い時間スケールのデータに関する解析である．もう一つは非線形力学系のモデルに対する統計科学の手法の適用である．

まず，タンパク質の分子動力学データの解析に関しては，Protein Data Bank（以下でPDBと略す）で1TIBとして登録されている酵素（正確には，TTLと呼ばれる酵素の安定構造の一つ）に対する時系列解析である（Kamada, Toda, Sekijima, Takada and Joe, accepted and to be published in Chemical Physics Letters）．時系列データの時間スケールは全長で２ナノ秒であり，タンパク質の機能に関して何か主張するには短すぎるが，解析手法のテストとして手始めに行うには良い．この研究では，ウエーブレット解析と特異値分解を用いた「集団運動」の抽出と，その「集団運動」相互の相関に関する解析を行った．

興味深いのは，この酵素の運動において，シミュレーション時間が短いにも関わらず，複数の安定構造に渡って動いていることを示唆する結果が出ていることである．これは，TTL が複数の安定構造を持つ（PDBには，1TIBと同じアミノ残基の並びを持つ酵素の安定構造が，全部で７つ登録されている）という特性に依ると考えられるが，この結果は，いわゆる「天然変性タンパク質」との関連でも重要であると考えられる．「天然変性タンパク質」ではタンパク質がただ一つの安定構造の周りに運動しているのではなく，より大きな変形運動を行っており，その運動の様相が分子機能にとって重要と考えられている．このように複数の安定構造を渡って運動する分子や，そもそも安定構造を持たない分子に対する解析において，我々の解析手法は新たな視点を提供できると期待している．特にウエーブレット変換を用いた解析は，時間的に変動する集団運動の解析に適しており，今後の展開が期待できる．

第二の点である統計科学の手法に依る非線形モデルの解析では，これまで共同研究者である福水による結果があるが，反応過程のモデルであるハミルトン力学系に対する解析は従来ない．本研究では，統計的反応論の基礎を与える相空間構造の不変集合，特に法双曲的不変多様体（Normally Hyperbolic Invariant Manifolds, NHIMsと略す）とその安定・不安定多様体（M. Toda, Global Aspects of Chemical Reactions in Multi-dimensional Phase Space, Advances in Chemical Physics, **130**, Part A, 337-399 (2005)）を時系列データから再構築する統計的手法の開拓，および，高次元の相空間構造を視角化するデータ解析手法の探求を行っている．

また，これら非線形科学の解析手法を小規模データに試験的に適用するため，共同研究者である藤崎とともに比較的小さな分子である N-methylacetamide の解析を進めている．この分子における振動緩和の時系列データに対するウエーブレット解析と，非線形共鳴の成す高次元の相空間構造の視角化が現時点での課題である．

特に興味があるのは，これらの振動運動における「遅い緩和」のメカニズムである．従来の統計的反応論では，これらの振動緩和は強い混合性のために指数関数的に生じると考えられている．しかし実験的には様々な分子で「遅い緩和」の存在が知られており，少数自由度のハミルトン系では「ベキ的緩和」を示すモデル系も知られている．これらの研究において欠けているのは，「中程度の自由度」の系における「遅い緩和」の動力学的メカニズムの解明である．この場合，「中程度の自由度」とは，たとえば数十程度の粒子から構成される系であり，N-methylacetamide はその典型例である．このような研究は，「遅い緩和」における量子効果の問題や，非線形共鳴を持つ系における反応制御の可能性，タンパク質などより大きな生体分子における「動力学的な非平衡性」の可能性など，本研究の様々な課題と関連する．またウエーブレットで抽出された特徴的時間スケールの異なる振動緩和モードにおいて，それらの間の独立性

解析を，福水によって展開されているカーネル法の手法を応用することで遂行している．

次に，検証作業が終わりデータ処理の手法が確立した段階で応用計算を進めるために，共同研究者の高見を中心に開発が進められている時系列データ解析システムについて述べる．本システムの公開時期については未定であるが，分子動力学等の数値計算時系列のデータ解析を扱う処理システムとして構築し，ネットワークを介して利用でき，バックエンドには必要に応じて九州大学の大規模計算機を利用できるように構成する．現時点での全体のシステム構成は図１のようになる．

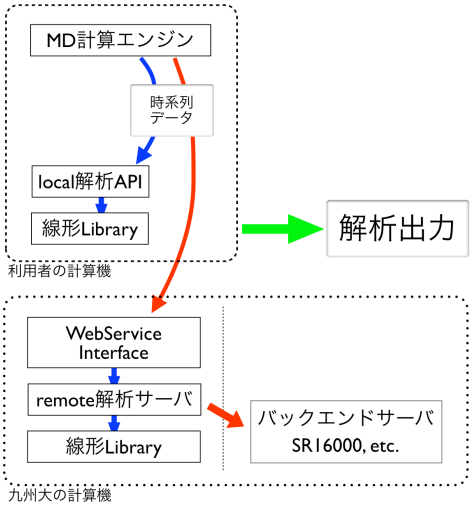

図１　解析プログラムの構造

本研究で開発される時系列解析の手法は，[local 解析 API]に実装されており，利用者側で分子動力学計算プログラム，あるいは，時系列処理プログラムから直接利用することが可能で，この場合は，利用者計算機上のプログラムとして実行される．しかし，計算が大規模になるなどの理由で利用者側の計算機だけでは利用が難しい場合には，Web Service のインターフェイスを通じてネットワーク越しに時系列データを送付し，九州大学で運用する[remote 解析サーバ]に解析を依頼することが可能である．この場合，必要に応じて，九州大学が運用する高性能計算機(SR16000)などをバックエンドとして解析計算を実行できるようになっている．

4.　これまでの進捗状況と今後の展望

まず最初に，本研究の実施に関連した我々の取り組みについて解説した上でこれまでの研究の経過を述べ，その上で，今後の研究計画と展望について記述する．

近年の一分子計測実験の発展は，生体高分子の運動の非定常性を明らかにしている．これを受けて，非平衡非定常な環境の下で生体分子の機能発現がいかに可能なのか，分子における階層的集団運動が動的に連関して生起する機構の解明が問われている．課題責任者の戸田は，大自由度カオス系に基づく非平衡反応動力学の構築[Adv. Chem. Phys. **130A**, 337 (2005)]，および，生体分子の分子動力学データから機能に寄与する集団運動抽出の試み(図 2)を行ってきた(櫻井・戸田・渕上・木寺，「タンパク質分子の分子動力学に対する時系列解析」，日本物理学会，2008 年～2010 年，および日本生物物理学会，2010 年)．

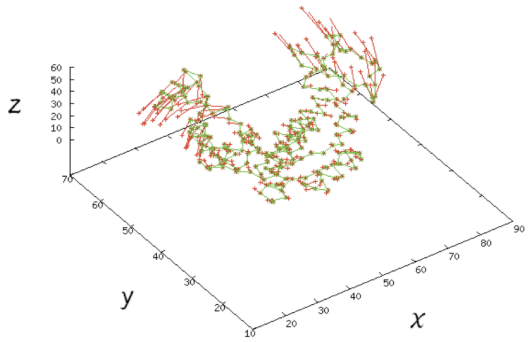

図２　時系列解析の結果

生体分子の機能発現では，特徴的な時空間スケールを持つ複数の階層で様々な集団運動が生成・崩壊・転生を繰り返す．これら階層的集団運動が，揺れ動く外界のもとで連関して励起される機構を解明するという長期にわたる研究の一環として本研究を位置づけ，(1)大規模分子動力学データから階層的集団自由度の動的挙動を抽出し，(2)閉じた微分方程式で記述される「決定論」ではない集団運動の動的挙動から，それら集団運動相互の統計的相関・因果関係を解析し，その成果に立って(3)特徴的な時空間スケールのより大きい「粗視化」

モデルを構築，階層的な計算スキームを開発して行くことを計画している．

そのために，次のような計画で研究を進めている．本研究では生体分子における階層的集団運動の動力学の解明に向けて，計算科学・統計科学・非線型物理の近年の成果を発展させ，その結果を具体的な分子動力学データに適用することで理論的な有効性を検証する．階層的集団運動の抽出に応用される「データマイニング」の方法には，ウエーブレット解析・非線形主成分解析・独立成分解析・カーネル法・特異値分解・ランダム行列理論などがあるが，これらの手法は単独で用いられることはあっても，系統的に組み合わされて応用される事例はまだほとんどない．さらに本研究では，統計科学の統計的因果推論，大自由度力学系の成果である不変集合解析など近年の成果をさらに発展させ，揺らぐ外界における生体分子の機能発現に対し，その頑健性の動力学的メカニズムを解明する．以上の研究は計算科学・統計科学・生物物理・非線型物理など広い関連分野の成果を結集する必要があると考える．

本研究の実施体制は，以下の通りである．階層的集団運動を抽出するデータマイニングを最終的な目的とした階層的集団運動の時系列データに対する統計的相関・因果推論は，計算科学の高見と統計科学の福水が協力して担当する．特に理論的な方法論の発展は統計科学の福水が担当し，アルゴリズムの実装は計算科学の高見が担当する．大自由度力学系の相空間構造・不変集合 による階層的集団運動の解析は，非線型物理の戸田が主に担当する．これらの成果を結集して，分子機能発現の頑健性の動力学的メカニズムを解明する．本研究の理論的方法論を実際に適用する分子動力学計算の時系列データは，大規模な検証では九州大学情報基盤研究開発センターの計算機(SR16000)を利用して，日本医科大学の藤崎氏・菊地氏が分担して実施する予定である．ただし以上の体制は流動的であり，相互的な乗入れが起ることが予想される．

これまでの研究の進捗については，以下の通りである．時系列データに関しては，九州大学の計算機を利用して大規模に取得することを計画しており，これまで最終的なデータの取得に向けて準備をしている段階である．分子動力学計算による時系列の取得については,Amber等のパッケージソフトウェアを利用して取得すること自体には，データ量の規模の問題をのぞいて困難な部分は生じないと考えられる．むしろ，データの解析手法の確立と検証に向けたサンプルデータの作成とデータ解析手法の比較検討に時間を要している．

今後の研究は，時系列データの高精度化と大規模化を行い，さらに解析手法の大規模な検証へと進めて行く予定である．まず，大規模分子の時系列取得までの手法は，以下の通りとなる．生体分子の時系列解析を行うために，実験・理論(計算)両面からよく調べられているタンパク質，ミオグロビンを取り上げる．ここで扱う時系列データは，大きく分けて次の二種類である．一つは，全原子の分子動力学データから得られる大自由度力学系の時系列データで，もう一つは，ミオグロビン分子を粗視化した形で得られるモデル化された比較的低自由度の時系列データである．まず，全体の分子から粗視化されたモデル系を構築する手順を解説する．分子の機能に密接に関わっているヘムとヒスチジンを含むミオグロビン内の領域を切り取り，この部分については九州大学の計算機上で利用可能な汎用電子状態計算パッケージGaussianによる DFT 計算を行い,残りの部分は精度を落として半経験的計算を利用する．そうして得られた近似的なポテンシャルを用いて基準振動解析を行い，基準モードや調和振動数を算出した上で，励起するモードと結合している熱浴モードを抜き出し，繰り込み的な手法に基づいた自由度の縮約を行う．これにより振動モードに基づいた比較的低自由度のモデル分子が構築される．このモデル分子に対しては，量子力学的なダイナミクスの計算も可能なため，外場による制御の可能性を探る研究を行う． 次に，このようにして得られたモデル系と全自由度を含む分子全体の動力学計算を，九

州大学のスーパーコンピュータ上で実施し，時系列データを取得する．カーネル法による統計的解析については，サンプルデータで検証された手法を，上記の大規模データに対して適用する形で検証を行い，時系列データに対する解析手法としての確立を目指す．

また，これまで試験的に作成してきた解析用のプログラムに関しては，今後，最終的には公開することを目指して，周辺機能を提供するソフトウェアとあわせて，時系列解析システムとして開発していく予定である．現時点では，その完成と公開の時期について明言は出来ないが，階層的な動力学データに対する解析システムとして開発を続けることとしている．

5. 研究成果リスト

(1) 学術論文（投稿中のものは「投稿中」と明記）

1. Analysis of motion features for molecular dynamics simulation of proteins, M. Kamada, M. Toda, M. Sekijima, M. Takada and K. Joe, accepted and to be published in Chemical Physics Letters.

2. A dynamical switching of a reaction coordinate to carry the system through to a different product state at high energies, H. Teramoto, M. Toda and T. Komatsuzaki, submitted.

3. Ergodic Problems for Real Complex Systems in Chemical Physics, T. Komatsuzaki, A. Baba, S. Kawai, M. Toda, J. E. Straub, R. S. Berry, in press in Advances in Chemical Physics.

4. Non-Brownian Phase Space Dynamics of Molecules, the Nature of their Vibrational States, and non-RRKM Kinetics, D. M. LEITNER, Y. MATSUNAGA, A.SHOJIGUCHI, C-B. LI, T. KOMATSUZAKI, and M. TODA, in press in Advances in Chemical Physics.

5. Dynamical Reaction Theory based on Geometric Structures in Phase Space, S. Kawai, H. Teramoto, C-B. Li, T. Komatsuzaki and M. Toda, in press in Advances in Chemical Physics.

6. Non-Markovian Theory of Vibrational Energy Relaxation and its Application to Biomolecular Systems, H. Fujisaki, Y. Zhang and J. E. Straub, in press in Advances in Chemical Physics.

7. Protein Functional Motions: Basic Concepts and Computational Methodologies, S. Fuchigami, Y. Matsunaga, H. Fujisaki and A. Kidera, in press in Advances in Chemical Physics.

8. H. Fujisaki, M. Shiga and A. Kidera, Onsager-Machlup action-based path sampling and its combination with replica exchange for diffusive and multiple pathways, J. Chem. Phys., **132**, 134101 (2010).

(2) 国際会議プロシーディングス

(3) 国際会議発表

1. M. Toda, Time Series Analysis using Wavelet toward Molecular Dynamics Simulation of Proteins, The 13th Slovenia-Japan Seminar on Nonlinear Science and Waseda AICS symposium on nonlinear and nonequilibrium phenomena in complex systems, 2010, Nov 4th-6th, Waseda University.

2. M. Toda, Phase Space Structures for Systems of Large Degrees of Freedom, Slovenia-Japan Seminar on Nonlinear Science (Kansai,2010), 2010 Nov. 8th-9th, Osaka Prefecture University.

3. K. Fuji, Time Series Analysis Using Wavelet for Molecular Dynamics Simulation of Proteins : A case for chignolin, the first designed Protein, Slovenia-Japan Seminar on Nonlinear Science (Kansai,2010), 2010 Nov 8th-9th, Osaka Prefecture University.

(4) 国内会議発表

1. 戸田, Dynamics in Complex Syetems (北大, 2011年3月7日-9日)で招待講演予定.

2. 高見 利也, 戸田 幹人, 福水 健次「分子動力学データに対する統計解析システムの構築と応用」HPCS2011 (産総研, 2011年1月18日-19日)にてポスター発表予定(P1-4).

(5) その他 (特許, プレス発表, 著書等)

1. 福水健次「カーネル法入門 ―正定値カーネルによるデータ解析―」(朝倉書店, 2010).