統計数理(2014)
第 62 巻 第 2 号 243–255

特集「生体高分子の揺らぎとダイナミクス
—シミュレーションと実験の統計解析—」
[研究詳解]

独立成分分析 tICA でタンパク質の複雑な運動を解きほぐす

渕上 壮太郎†

(受付 2014 年 1 月 1 日；改訂 9 月 3 日；採択 9 月 8 日)

要　　旨

近年の計算機の目覚ましい発展にともない，タンパク質の運動を分子動力学シミュレーションで再現することができるようになった．しかし，タンパク質は多くの原子から構成される大自由度系であるため，その運動は複雑多様であり，運動の実態を把握・理解することは容易でない．シミュレーション結果からタンパク質の主要な運動を特定・抽出するために，様々なデータ解析手法が提案・開発・適用されてきたが，タンパク質の運動を詳細に理解できるようになったとは言い難く，さらなる方法の開発，研究の発展が強く望まれる．我々は，タンパク質の「遅い運動」に着目し，シミュレーション結果から効率的に同定するための手法として，「時間構造に基づいた独立成分分析(tICA)」を提案し，実際，この解析手法が有用であることを示した．本稿では，この tICA について，その定式化から実践までを詳しく解説するとともに，タンパク質主鎖の運動に適用した結果を紹介する．

キーワード：タンパク質ダイナミクス，分子動力学シミュレーション，独立成分分析，tICA，遅い運動，レアイベント．

1. はじめに

タンパク質は，生物のからだをつくり，動かしている基本的な分子であり，生命の仕組みを解き明かすためには，その構造・機能を理解することが必要不可欠である．また，タンパク質の異常や，細菌・ウイルス特有のタンパク質の働きを知ることで，病気の原因特定や効果的な治療薬の開発も期待できる．タンパク質は 20 種類のアミノ酸が一列につながった鎖状の高分子であり，アミノ酸の並びにしたがってそれぞれ固有の形へと折り畳まれる．このように，特定の立体構造を形成することによって，タンパク質はその機能を発揮することができる．したがって，タンパク質の働きやそのメカニズムを理解するためにはタンパク質の立体構造を明らかにすることが重要である．

しかし，タンパク質の立体構造だけで，その機能がすべて理解できるわけではない．タンパク質は柔軟性に富んだ動的な分子であり，生体内環境において大きく揺らいでいる．タンパク質は多くの原子から構成される大自由度系であるため，その揺らぎは幅広い時間・空間スケールに渡った複雑なものであることは想像に難くない．実際，タンパク質の揺らぎには，原子間結合の振動にはじまり，メチル基の回転，ループ運動，ドメイン運動まで様々な階層の運動が含まれており，その時間スケールはフェムト秒からミリ秒まで幅広い領域に渡っている．この

†横浜市立大学大学院 生命医科学研究科：〒 230–0045 横浜市鶴見区末広町 1–7–29

ようなタンパク質の揺らぎは，一見，タンパク質の機能実現を妨害する邪魔者に過ぎないように思えるが，実際にはそれほど単純ではない．

これまで，タンパク質が示す複雑多岐な運動について，実験・理論を問わず，様々な手法を用いた多角的な解析が行われ，機能との密接な関連が明らかになってきた（Henzler-Wildman and Kern, 2007; Fuchigami et al., 2011）．タンパク質は自身の揺らぎを巧みに制御し，さらには有効に活用することによって，高精度で高効率な機能を実現しているようである．したがって，タンパク質の機能を解明するためには，立体構造という静的な特徴だけでなく，タンパク質の動的側面，つまり「タンパク質ダイナミクス」をも理解する必要がある．タンパク質が機能を発現する時間スケールは一般的にマイクロ秒以上であることから，タンパク質の運動の中でも，ドメイン運動のような遅い時間スケールの運動が機能実現に関わっている可能性が高いと予想される．

近年のコンピュータおよびソフトウェアの劇的な発展により，分子動力学（MD）シミュレーションによって，タンパク質の運動をマイクロ秒以上の長時間に渡って再現することができるようになった．MD シミュレーションは，タンパク質の平衡揺らぎや立体構造変化，折り畳み過程などタンパク質が示す動的な挙動を詳細に調べる上で有力な手段であり，タンパク質の機能が実現されるメカニズムを原子レベルで明らかにすべく広く利用されている（Klepeis et al., 2009; Dror et al., 2012; Lane et al., 2013）．ただし，シミュレーションを実行するだけで，タンパク質の機能が即解明できるわけではないことは言うまでもない．MD シミュレーションでは，すべての原子の動きを追跡することができることから，得られるデータ量は膨大であるが，タンパク質機能の理解にとって欠かすことができない情報はそのごく一部で，大半の部分は不要であると考えられる．したがって，大自由度系であるタンパク質が示す高次元空間内の複雑な運動の中から，機能と関連する可能性が高い運動を特定し，抽出しなければならない．

MD シミュレーションで得られた結果を解析する手法はこれまでに数多く提案されているが，中でも最も広く使われているのが主成分分析（PCA）である（Hayward and Go, 1995; Kitao and Go, 1999; Berendsen and Hayward, 2000; Fuchigami et al., 2011）．PCA で得られるモード（主成分）はタンパク質の大振幅運動を表し，タンパク質が機能を発現する際によく見られる大規模な立体構造変化を少数の主成分でうまく表現できることが多くの研究で報告されている．しかし，PCA による解析だけでタンパク質の動的な振る舞いを十分に理解することは難しい．たとえば，PCA で抽出される大振幅モードは時間スケールの遅い運動を表していることが多いが，逆に，遅い運動だからといって大振幅運動であるとは限らず，このような運動が存在する場合，PCA による解析ではその存在を見逃してしまう可能性が高い．また，主成分が記述する運動は互いに無相関であるが，その独立性は保証されておらず，相互に関連してしまっていることが多い．したがって，各主成分方向の運動を個別に解析し，その重ね合わせとしてタンパク質全体の運動を理解しようとすることには問題がある．このような PCA の欠点を克服するため，これまでに様々な解析手法が提案・適用されてきたが，タンパク質ダイナミクスの全貌はいまだよくわかっていない．特に，タンパク質が示す遅い運動に関する解析が不十分と思われる．

そこで我々は，シミュレーションの結果からタンパク質の遅い運動を特定・抽出するための方法論を開発・検証し，その確立を目指した（Naritomi and Fuchigami, 2011; 渕上, 2011; Naritomi and Fuchigami, 2013）．解析手法として着目したのは，独立成分分析（ICA）の一種で，Molgedey と Schuster によって提案されたシステムの動的な特性を利用したアルゴリズム（Molgedey and Schuster, 1994）であり，開発した手法を「時間構造に基づいた独立成分分析（tICA）」と名付けた．本稿では，この tICA についてその原理から実践までを詳しく解説するとともに，具体例として，タンパク質主鎖の運動に適用した結果を紹介する．

2. 時間構造に基づいた独立成分分析(tICA)

ここでは，タンパク質の遅い運動を解明するための手法として我々が提案した「時間構造に基づいた独立成分分析(tICA)」について，その定式化を示すとともに，得られた結果の解釈，および，利用方法を説明する．

2.1 tICA の定式化

今，解析の対象となる n 次元の時系列データを $\boldsymbol{x}(t) = {}^t(x_1(t), x_2(t), \ldots, x_n(t))$ で表そう．ここで，左肩の t は転置を意味し，n 次元データを縦ベクトルで表現している．この時系列データ $\boldsymbol{x}(t)$ によって記述される n 次元空間における運動を，n 個の 1 次元運動に分解することを考えよう．各 1 次元運動の方向を示すベクトルを $\boldsymbol{g}_i$ $(i = 1, 2, \ldots, n)$ とすると，n 次元時系列データ $\boldsymbol{x}(t)$ は 1 次元運動の重ね合わせとして以下のように書き表すことができる：

$$\boldsymbol{x}(t) = a_1(t)\boldsymbol{g}_1 + a_2(t)\boldsymbol{g}_2 + \cdots + a_n(t)\boldsymbol{g}_n = \mathbf{G}\boldsymbol{a}(t). \tag{2.1}$$

ここで，$a_i(t)$ は $\boldsymbol{g}_i$ 方向の 1 次元運動を表す時系列である．また，$\mathbf{G}$ はベクトル $\boldsymbol{g}_i$ を並べてできる n 次元正方行列

$$\mathbf{G} := (\boldsymbol{g}_1\, \boldsymbol{g}_2\, \cdots\, \boldsymbol{g}_n), \tag{2.2}$$

であり，$\boldsymbol{a}(t)$ は $a_i(t)$ を要素とする n 次元時系列

$$\boldsymbol{a}(t) := {}^t(a_1(t), a_2(t), \ldots, a_n(t)), \tag{2.3}$$

である．このように，多次元運動を 1 次元運動へと分解することができれば，$a_i(t)$ によって表される各 1 次元運動を個別に解析することで全体の運動を把握・理解することができる．

式(2.1)のような運動の分解は，n 個の互いに独立なベクトルを用意し，それらを基底ベクトルとすれば常に可能である．しかし，運動の分解が意味を持つためには，n 個の基底ベクトルを適切に選ぶ必要があるだろう．では，どのように選ぶのが良いだろうか？ここで，運動を分解するということはどういうことなのかを改めて考えてみよう．分解によって得られた 2 つの 1 次元運動を比べたとき，互いに良く似た挙動を示していたとする．この場合，一方の運動を調べることによってもう一方の運動に関する情報を得ることができるので，運動がうまく分解できているとは言えないであろう．つまり，分解された個々の運動は，他の運動とは異なるその運動固有の特徴を持ち，互いに異なる挙動を示すこと，つまり，「互いに独立であること」が望まれる．

多変量データの解析に幅広く利用されている独立成分分析(ICA)では，各 1 次元運動を確率過程とみなし，それらが互いに統計的に独立となるような独立成分 $\boldsymbol{g}_i(i = 1, 2, \ldots, n)$ を見つけ出すのが一般的であり，様々なアルゴリズムが提案されている(Hyvärinen et al., 2001; ビバリネン 他, 2005; 甘利 他, 2002)．各確率過程がガウス分布に従う場合には，互いに無相関となるように独立成分を決めればよく，実際，そのような成分は主成分分析(PCA)によって簡単に見つけることができる．しかし，現実のほとんどの場合では，分布はガウス分布とならず，無相関性だけでは独立成分を決定するのに十分ではない．このことは，PCA で運動をうまく分解できない理由でもある．そこで，典型的な ICA では，分布の無相関性に加えて，「非線形変換を施したものの無相関化」や「非ガウス性の最大化」などの条件を課して独立成分の推定を行う．つまり，典型的な ICA では，無相関化に必要な二次の統計量に加えて，何らかの高次の統計量も用いることになる．

一方，各 1 次元運動が時間相関を持つ場合には，高次の統計量の代わりに，時間的な構造の

情報を利用することで，独立成分を推定することができる．我々が提案した tICA はこのような手法のひとつであり，時間構造の情報を用いることがその名前の由来となっている．具体的に仮定する条件として最も簡単なものは，分解で得られた 2 つの 1 次元運動 $a_i(t)$ と $a_j(t)$（ただし，$i \neq j$）の相互共分散関数 $c_{ij}(s)$ が任意の時刻 s においてゼロとなることである：

$$
\begin{aligned}
c_{ij}^{\boldsymbol{a}}(s) &= \langle (a_i(t) - \langle a_i(t) \rangle)(a_j(t+s) - \langle a_j(t) \rangle) \rangle \\
&\equiv 0.
\end{aligned} \tag{2.4}
$$

ここで，$\langle \cdots \rangle$ は時間平均を意味する．すべての i と j の組み合わせに対する条件をまとめて行列表示すると，以下のように書くことができる：

$$
\begin{aligned}
\mathbf{C}^{\boldsymbol{a}}(s) &= \left\langle (\boldsymbol{a}(t) - \langle \boldsymbol{a}(t) \rangle)\, {}^t(\boldsymbol{a}(t+s) - \langle \boldsymbol{a}(t) \rangle) \right\rangle \\
&= \mathrm{diag}(c_{11}^{\boldsymbol{a}}(s), c_{22}^{\boldsymbol{a}}(s), \ldots, c_{nn}^{\boldsymbol{a}}(s)).
\end{aligned} \tag{2.5}
$$

ここで，最左辺の行列 $\mathbf{C}^{\boldsymbol{a}}(s)$ は時系列 $\boldsymbol{a}(t)$ の相互共分散関数行列であり，その対角要素である $c_{ii}^{\boldsymbol{a}}(s)$ は各運動 $a_i(t)$ の自己共分散関数である．$\mathbf{C}^{\boldsymbol{a}}(s)$ は時刻 s の関数であるが，ある特定の値（たとえば $s = t_0$）を代入した $\mathbf{C}^{\boldsymbol{a}}(t_0)$ は時間遅れ共分散行列，もしくは，時間差共分散行列と呼ばれる．特に，$s = 0$ の場合，$\mathbf{C}^{\boldsymbol{a}}(0)$ は $\boldsymbol{a}(t)$ の共分散行列である．この等式に，左から $\mathbf{G}$ を，右から ${}^t\mathbf{G}$ をそれぞれ掛けると，

$$
\begin{aligned}
\mathbf{G}\mathbf{C}^{\boldsymbol{a}}(s)\,{}^t\mathbf{G} &= \mathbf{G} \left\langle (\boldsymbol{a}(t) - \langle \boldsymbol{a}(t) \rangle)\, {}^t(\boldsymbol{a}(t+s) - \langle \boldsymbol{a}(t) \rangle) \right\rangle {}^t\mathbf{G} \\
&= \left\langle (\mathbf{G}\boldsymbol{a}(t) - \langle \mathbf{G}\boldsymbol{a}(t) \rangle)\, {}^t(\mathbf{G}\boldsymbol{a}(t+s) - \langle \mathbf{G}\boldsymbol{a}(t) \rangle) \right\rangle \\
&= \left\langle (\boldsymbol{x}(t) - \langle \boldsymbol{x}(t) \rangle)\, {}^t(\boldsymbol{x}(t+s) - \langle \boldsymbol{x}(t) \rangle) \right\rangle \\
&= \mathbf{C}(s),
\end{aligned} \tag{2.6}
$$

となる．ここで，時系列データ $\boldsymbol{x}(t)$ の相互共分散関数行列を $\mathbf{C}(s)$ とした．これより，独立成分を推定するには，時系列データ $\boldsymbol{x}(t)$ の相互共分散関数行列 $\mathbf{C}(s)$ を任意の時刻 s において対角にするベクトルを求めれば良いことがわかる．

実際には，2 つの時刻においてこの等式が成り立つことを要請すれば，独立成分の行列 $\mathbf{G}$ を決定することができる．tICA では，2 つの時刻として $s = 0$ と $s = t_0$ を用いる．したがって，tICA で解くべき問題は $\mathbf{C}(0)$ と $\mathbf{C}(t_0)$ の同時対角化となる：

$$
\begin{cases}
\mathbf{C}(0) = \mathbf{G}\mathbf{C}^{\boldsymbol{a}}(0)\,{}^t\mathbf{G}, \\
\mathbf{C}(t_0) = \mathbf{G}\mathbf{C}^{\boldsymbol{a}}(t_0)\,{}^t\mathbf{G}.
\end{cases} \tag{2.7}
$$

ここで，2 つの式はそれぞれ $\mathbf{C}(0)$ と $\mathbf{C}(t_0)$ の固有値問題となっているが，それぞれの対角化によって得られる 2 つの固有ベクトルは一般に一致しない．したがって，同時固有ベクトルの行列 $\mathbf{G}$ を求めるには，個々の固有値問題を解くのではなく，2 つの式から導出される以下の一般化固有値問題

$$
\mathbf{C}(t_0)\mathbf{F} = \mathbf{C}(0)\mathbf{F}\mathbf{K} \tag{2.8}
$$

を解かなければならない．ここで，$\mathbf{K}$ と $\mathbf{F}$ がそれぞれ一般化固有値問題の固有値行列と固有ベクトル行列である．固有値行列 $\mathbf{K}$ は $\mathbf{K} = \mathbf{C}^{\boldsymbol{a}}(t_0)\mathbf{C}^{\boldsymbol{a}}(0)^{-1}$ と表される．また，求めるべき行列 $\mathbf{G}$ は，得られた固有値ベクトル行列 $\mathbf{F}$ を用いて $\mathbf{G} = \mathbf{C}(0)\mathbf{F}$ で与えられる．以上によって，tICA による解析を定式化することができた．

2.2 tICA を用いた運動の分解

tICA の定式化ができたので，続いて，n 次元時系列データ $\boldsymbol{x}(t)$ で表される多次元空間の運動を tICA を用いて 1 次元運動に分解してみよう．tICA を実行するには，共分散行列 $\mathbf{C} := \mathbf{C}(0)$ と時間遅れ共分散行列 $\bar{\mathbf{C}} := \mathbf{C}(t_0)$ が必要であった：

$$\mathbf{C} = \left\langle (\boldsymbol{x}(t) - \langle \boldsymbol{x}(t) \rangle)\, {}^t(\boldsymbol{x}(t) - \langle \boldsymbol{x}(t) \rangle) \right\rangle, \tag{2.9}$$

$$\bar{\mathbf{C}} = \left\langle (\boldsymbol{x}(t) - \langle \boldsymbol{x}(t) \rangle)\, {}^t(\boldsymbol{x}(t+t_0) - \langle \boldsymbol{x}(t) \rangle) \right\rangle. \tag{2.10}$$

ここで，t_0 は遅延時間パラメータであり，その値は $\boldsymbol{x}(t)$ で記述される運動の時間スケールに合わせて適切に決定する必要がある．これら 2 つの行列を用いて，前節で定式化された一般化固有値問題

$$\bar{\mathbf{C}}\mathbf{F} = \mathbf{C}\mathbf{F}\mathbf{K}, \tag{2.11}$$

を解き，固有値行列 $\mathbf{K} = \mathrm{diag}(k_1, k_2, \ldots, k_n)$ と固有ベクトル行列 $\mathbf{F} = (\boldsymbol{f}_1\ \boldsymbol{f}_2\ \cdots \boldsymbol{f}_n)$ を求める．一般的に，$\bar{\mathbf{C}}$ は非対称行列なので，一般化固有値問題(2.11)の固有値や固有ベクトルの要素は複素数となる．複素数を避けるためには，$\bar{\mathbf{C}}$ を対称化した行列 $\bar{\mathbf{C}}_{\mathrm{sym}} = (\bar{\mathbf{C}} + {}^t\bar{\mathbf{C}})/2$ を $\bar{\mathbf{C}}$ の代わりに用いればよい．この対称化は時系列が時間反転に関して対称であれば正当化され，実際，多くの場合，この仮定は満たされている．

tICA で得られる固有ベクトル $\boldsymbol{f}_i$ は，PCA の主成分と異なり，互いに直交せず，直交基底を成していない．しかし，$\boldsymbol{f}_i$ は互いに独立なので，非直交基底として利用することは可能である．$\boldsymbol{f}_i$ は関係式

$${}^t\boldsymbol{f}_i \mathbf{C} \boldsymbol{f}_j = \delta_{ij}, \tag{2.12}$$

を満たすように決定できるので，$\boldsymbol{f}_i$ の双対ベクトル $\boldsymbol{g}_i$ を

$$\boldsymbol{g}_i = \mathbf{C}\boldsymbol{f}_i, \tag{2.13}$$

と定義すれば，以下の関係が成り立つ：

$${}^t\boldsymbol{f}_i \boldsymbol{g}_i = \delta_{ij}, \tag{2.14}$$

$$\sum_i \boldsymbol{g}_i\, {}^t\boldsymbol{f}_i = \mathbf{1}. \tag{2.15}$$

ここで，$\mathbf{1}$ は n 次元の恒等行列である．したがって，この非直交基底を用いると，時系列データ $\boldsymbol{x}(t)$ を以下のように展開することができる：

$$\boldsymbol{x}(t) = \sum_i \boldsymbol{g}_i\, {}^t\boldsymbol{f}_i \boldsymbol{x}(t) = \sum_i a_i(t) \boldsymbol{g}_i. \tag{2.16}$$

ここで，$a_i(t) = {}^t\boldsymbol{f}_i \boldsymbol{x}(t)$ は双対ベクトル $\boldsymbol{g}_i$ の方向の 1 次元運動を表す時系列である．この式は，tICA の定式化の最初で，n 次元時系列データ $\boldsymbol{x}(t)$ を 1 次元運動の重ね合わせとして書き表した式(2.1)と同じ表式である．つまり，tICA では，独立な運動の方向を表す独立成分は，固有ベクトル $\boldsymbol{f}_i$ ではなく，その対となる $\boldsymbol{g}_i$ であることがわかる．また，独立成分へ時系列データ $\boldsymbol{x}(t)$ を射影する際には，$\boldsymbol{g}_i$ ではなく，$\boldsymbol{f}_i$ を使わなければならない．以上のように，tICA で得られた固有ベクトルを用いることで，n 次元時系列データ $\boldsymbol{x}(t)$ を 1 次元運動 $\boldsymbol{a}(t) = {}^t(a_1(t), a_2(t), \ldots, a_n(t)) = {}^t\mathbf{F}\boldsymbol{x}(t)$ に分解できることがわかった．

では，分解で得られた 1 次元運動は，必要な性質を備えているだろうか？ $\boldsymbol{a}(t)$ に求められる性質は以下の 2 つである：

(1) $\boldsymbol{a}(t)$ の共分散行列は対角行列となる．
(2) $\boldsymbol{a}(t)$ の遅延時間 t_0 の時間遅れ共分散行列は対角行列となる．

まず，一つ目の性質から確認してみよう．$\boldsymbol{a}(t)$ の共分散行列を計算すると以下のようになる：

$$\begin{aligned}\left\langle (\boldsymbol{a}(t)-\langle \boldsymbol{a}(t)\rangle)^t(\boldsymbol{a}(t)-\langle \boldsymbol{a}(t)\rangle)\right\rangle &= \left\langle ({}^t\mathbf{F}\boldsymbol{x}(t)-\left\langle {}^t\mathbf{F}\boldsymbol{x}(t)\right\rangle)\,{}^t({}^t\mathbf{F}\boldsymbol{x}(t)-\left\langle {}^t\mathbf{F}\boldsymbol{x}(t)\right\rangle)\right\rangle \\ &= {}^t\mathbf{F}\left\langle (\boldsymbol{x}(t)-\langle \boldsymbol{x}(t)\rangle)\,{}^t(\boldsymbol{x}(t)-\langle \boldsymbol{x}(t)\rangle)\right\rangle \mathbf{F} \\ &= {}^t\mathbf{F}\mathbf{C}\mathbf{F}=\mathbf{1}.\end{aligned} \tag{2.17}$$

これより，$\boldsymbol{a}(t)$ が確かに一つ目の性質を満たしていることがわかる．続いて，遅延時間 t_0 の時間遅れ共分散行列も計算してみると，

$$\begin{aligned}\left\langle (\boldsymbol{a}(t)-\langle \boldsymbol{a}(t)\rangle)^t(\boldsymbol{a}(t+t_0)-\langle \boldsymbol{a}(t)\rangle)\right\rangle &= \left\langle ({}^t\mathbf{F}\boldsymbol{x}(t)-\left\langle {}^t\mathbf{F}\boldsymbol{x}(t)\right\rangle)\,{}^t({}^t\mathbf{F}\boldsymbol{x}(t+t_0)-\left\langle {}^t\mathbf{F}\boldsymbol{x}(t)\right\rangle)\right\rangle \\ &= {}^t\mathbf{F}\left\langle (\boldsymbol{x}(t)-\langle \boldsymbol{x}(t)\rangle)\,{}^t(\boldsymbol{x}(t+t_0)-\langle \boldsymbol{x}(t)\rangle)\right\rangle \mathbf{F} \\ &= {}^t\mathbf{F}\bar{\mathbf{C}}\mathbf{F}={}^t\mathbf{F}\mathbf{C}\mathbf{F}\mathbf{K}=\mathbf{K},\end{aligned} \tag{2.18}$$

となり，二つ目の性質も確かめられた．以上のように，tICA による運動の分解は期待通りの結果を与えることがわかる．

上式(2.18)より，$a_i(t)$ の自己相関関数は，時刻 $s=t_0$ において tICA の固有値 k_i に一致することがわかる．よって，固有値の値が大きいほど，自己相関関数の緩和が遅いということになり，対応する独立成分が表す運動の時間スケールが遅い，と考えることができる．もし $a_i(t)$ の自己相関関数が指数関数的な緩和を示すのであれば，その時定数(もしくは平均寿命)は

$$\tau_i=-\frac{t_0}{\ln k_i}, \tag{2.19}$$

と表される．一般に，自己相関関数は指数関数的な緩和を示すとは限らないが，多くの場合，固有値 k_i を用いて定義される上式(2.19)の τ_i が運動の時間スケールの目安となることが期待される．以上のように，tICA の固有値は対応する独立成分が表す運動の時間スケールを特徴づけており，最大の固有値をもつ第一独立成分(IC1) $g_1(t)$ が最も遅い運動の方向となることがわかる．これより，大きな固有値に対応する独立成分に着目すれば，時系列データが表す多次元空間内の多様な運動の中から遅い時間スケールをもった運動のみを抽出し，その挙動を調べることができる．

3. tICA の応用例：タンパク質主鎖が示す遅い運動

タンパク質の運動は複雑多様であり，独立な 1 次元運動の重ね合わせとして記述できるものではない．したがって，独立成分分析の前提が破綻しているように思われるが，タンパク質の運動を独立成分の重ね合わせとして近似することは可能である．そのような近似的独立成分は tICA によって特定することができ，特に大きな固有値をもつ独立成分はタンパク質の遅い運動を表していると期待される．この節では，tICA を用いてタンパク質の運動を解析した例を紹介する(Naritomi and Fuchigami, 2013)．

解析の対象としたタンパク質は「リジン・アルギニン・オルニチン結合タンパク質(LAO)」である．LAO は 238 残基，3,649 原子からなり，図 1(a)に示したように，2 つのドメインをもつ．ドメインの境界部分には大きな裂け目が存在し，そこにリジンやアルギニン，オルニチンといったリガンドが特異的に結合する．このリガンドの結合にともない，LAO は大きな立体構造変化を起こすこともわかっている(Oh et al., 1993, 1994)．また，リガンドが結合していない場合，LAO は大きなドメイン揺らぎを示すことが予想されるとともに，このドメイン揺らぎがリガン

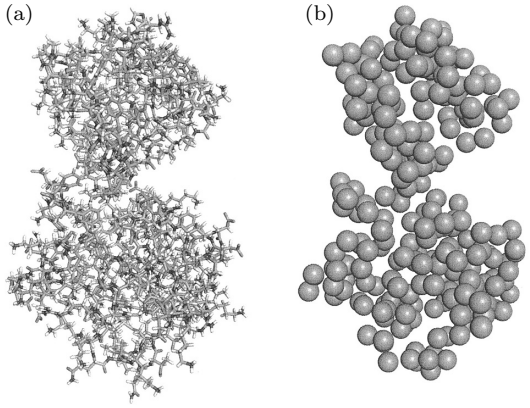

図 1．LAO の立体構造．左：全原子モデル．右：C_α 原子のみを表示．

ド結合時の立体構造変化に活用されていると考えられる．

そこで，リガンド非結合時の LAO が示す揺らぎを明らかにすべく，LAO を大量の水分子（25,392 分子）の中に埋め込み，水中の状態を再現した系（総原子数 79,828 原子）を構築し，1 μs の長時間 MD シミュレーションを実行した．シミュレーションの実行には，池口によって開発された分子シミュレーションプログラム MARBLE（Ikeguchi, 2004）を使用し，力場には CHARMM22/CMAP（MacKerell et al., 1998, 2004）を用いた．計算の詳細については原著論文（Naritomi and Fuchigami, 2013）を参照してほしい．

3.1 tICA による LAO 主鎖のダイナミクス解析

上述のように LAO は 3,649 原子からなるタンパク質であり，その自由度は 10,947 ($= 3,649 \times 3$) にものぼる．これらの全自由度を対象とした解析を行うことも可能であるが，ここではタンパク質主鎖の運動に着目し，C_α 原子のみを対象として tICA による解析を行った．これは，タンパク質の遅い運動は，そのほとんどが主鎖の運動に起因すると考えられ，その運動は C_α 原子だけで十分に表現され得ると思われるからである（図 1（b）を参照）．これにより，解析すべき時系列データの次元は 714 ($= 238 \times 3$) と大幅に減少し，解析が容易となる．実際には，C_α 原子の位置座標の時系列データそのものに tICA を適用するのではなく，外部自由度に由来する問題を回避するため，少々複雑な手順を用いて時系列データを変換した後に解析を行ったのであるが，その詳細はここでは割愛する．

C_α 原子を対象とした tICA による解析から，LAO 主鎖の遅い運動としてどのようなものが特定されたのか見てみよう．ここで，tICA のパラメータである遅延時間 t_0 は 1.0 ns とした．まず，tICA で得られた独立成分のうち，固有値が大きいもの 5 つについて，その運動の時間スケールを確認してみたところ，式（2.19）で定義される時定数 τ_i ($i = 1, 2, \ldots, 5$) はそれぞれ IC1：28.0 ns，IC2：13.4 ns，IC3：10.7 ns，IC4：6.6 ns，IC5：4.5 ns であった．これより，tICA で得られた上位の独立成分で表される運動は，数十ナノ秒から数ナノ秒の時間スケールをもった遅い運動であることがわかる．では，これらの遅い運動はどのような運動であろうか？

図 2（a）–（e）に示された上位 5 つの IC が表す運動を見てみると．IC2 が典型的なドメイン運動であるのに対し，それ以外の 4 つ（IC1，IC3，IC4，IC5）では著しい動きを示す C_α 原子があることから，それらの近辺で局所的な運動が生じていると考えられる．この運動の局所性は，各

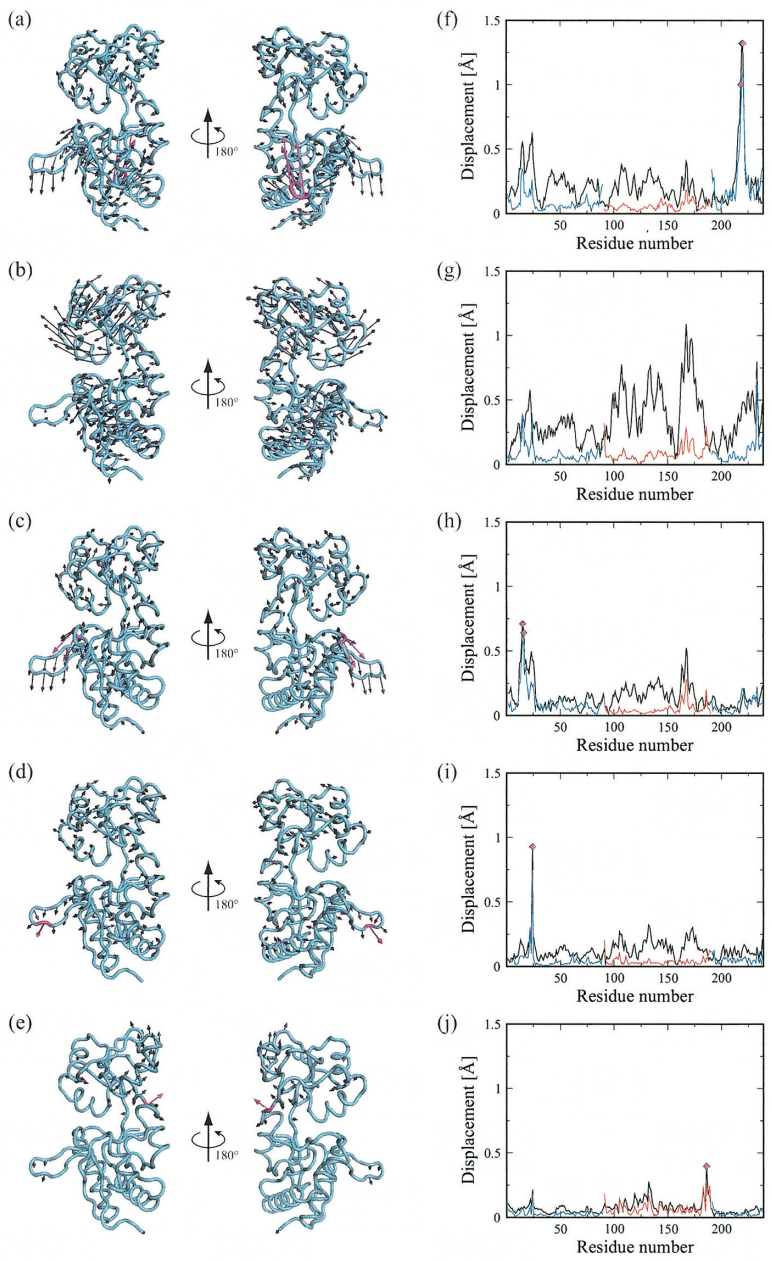

図 2. tICA によって特定された LAO 主鎖の遅い運動．(a)–(e) IC1 から IC5 によって表される C_α 原子の運動を矢印で示した．特に，顕著な変位を示す C_α 原子の矢印は紫色とした．(f)–(j) IC1 から IC5 によって誘起される C_α 原子の変位(黒線)．赤線と青線はドメイン運動を取り除いた場合の変位．紫色のダイアモンドは著しく動いている C_α 原子を示す．Reprinted with permission from Naritomi and Fuchigami (2013), J. Chem. Phys. 139, 215102. Copyright 2013 AIP Publishing LLC.

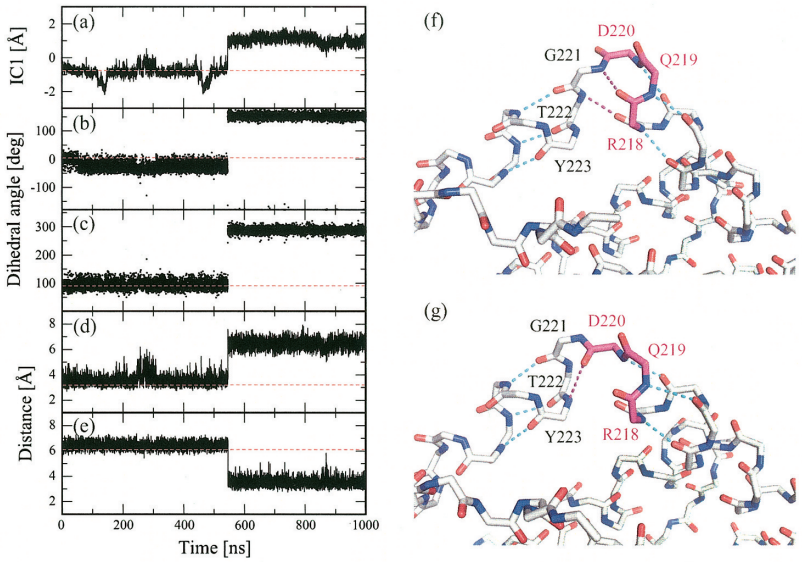

図 3. IC1 によって特定された LAO 主鎖の局所運動．(a)–(e)それぞれ，IC1，主鎖二面角 D220ψ，主鎖二面角 G221ϕ，原子間距離 R218O–G221N，原子間距離 D220O–Y223N の時間発展．シミュレーションの初期構造に用いた結晶構造の値を赤の破線で示した．(f)–(g)著しい動きを示す C_α 原子周辺の結晶構造，および，1 μ s 後の最終構造における LAO 主鎖．顕著な変位を示す C_α 原子を含む残基の主鎖炭素原子を紫色で示した．また，水色と紫色の破線は，安定，および，不安定な主鎖間の水素結合を表している．Reprinted with permission from Naritomi and Fuchigami (2013), J. Chem. Phys. 139, 215102. Copyright 2013 AIP Publishing LLC.

IC における C_α 原子の変位(図 2(f), (h)–(j))においてごく少数の大きな変位が存在することからも確認することができる．これらの顕著な変位はドメイン運動に由来するアーティファクトの可能性も考えられるが，IC が表す運動からドメイン運動の寄与を取り除いてもごく少数の大きな変位が維持されることから，確かに局所的な運動であると結論づけることができる．しかし，これらの局所運動の詳細を明らかにするには更なる解析が必要である．以下では，IC1，および，IC3 で記述される LAO 主鎖の局所運動を解析した結果を紹介する．

3.2 IC1 によって特定された LAO 主鎖の遅い局所運動

最も遅い運動を表している IC1 では，3 つの残基(R218，Q219，D220)において特に顕著な変位が見られ，その周辺で局所的な運動が生じていることがわかった．これら 3 つの残基は，結晶構造において，α ヘリックスの末端に位置し，複数の水素結合によって安定な構造を形成している(図 3(f))．それにもかかわらず，tICA の結果は，この局所部分に有意な構造変化が起こったことを示唆している．では，いったいどんなタイミングでどのような変化が起こったのだろうか？

図 3(a)に示した IC1 の時間発展からは，この部分の立体構造が 545.2 ns で遷移的に変化し，その後シミュレーションが終わるまで遷移後の構造のまま安定でいたことが見て取れる．実際，結晶構造(図 3(f))と 1 μs 後の立体構造(図 3(g))とを比較すると，残基 D220 と G221 の間のペ

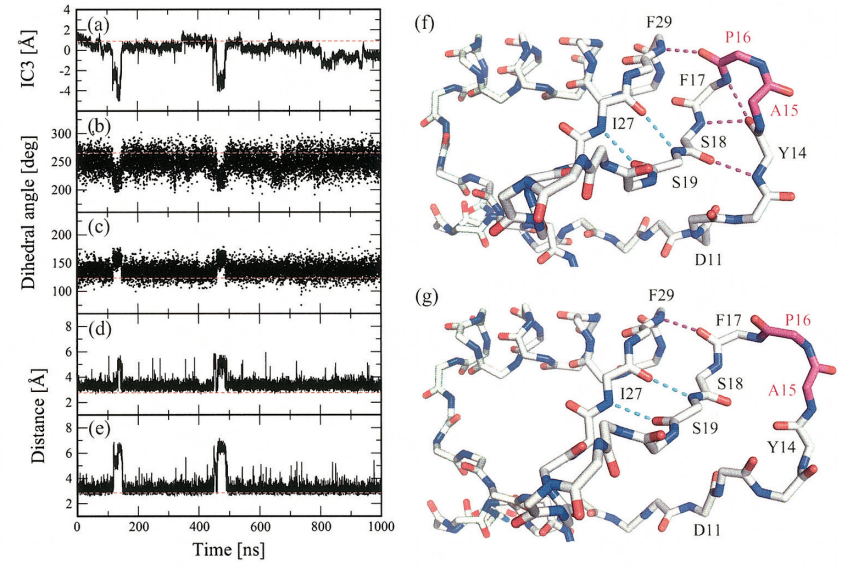

図 4. IC3 によって特定された LAO 主鎖の局所運動．(a)–(e)それぞれ，IC3，主鎖二面角 D11ϕ，主鎖二面角 S18ψ，原子間距離 Y14O–F17N，原子間距離 P16O–F29N の時間発展．シミュレーションの初期構造に用いた結晶構造の値を赤の破線で示した．(f)–(g)著しい動きを示す C_α 原子周辺の結晶構造，および，470 ns 時の構造における LAO 主鎖．顕著な変位を示す C_α 原子を含む残基の主鎖炭素原子を紫色で示した．また，水色と紫色の破線は，安定，および，不安定な主鎖間の水素結合を表している．Reprinted with permission from Naritomi and Fuchigami (2013), J. Chem. Phys. 139, 215102. Copyright 2013 AIP Publishing LLC.

プチド結合部分がクランクシャフト運動を起こすことで局所的に構造が変化している一方，近接する部分にはほとんど影響が及んでいないことがわかる．また，この運動に伴って 2 つの水素結合 R218O–G221N と D217O–T222N（O と N はそれぞれ主鎖の酸素原子と窒素原子を意味する）が切断され，新たに D220O と Y223N の間に水素結合が形成されたこともわかる．クランクシャフト運動の遷移的な挙動は，主鎖二面角 D220ψ と G221ϕ の遷移や，関連した水素結合の形成・切断が，IC1 と同一時刻でただ 1 回のみ起こっていることから確認することができる（図 3(b)–(e)）．以上より，IC1 で示唆された局所運動は確かに起こっており，その詳細を原子レベルで明らかにすることができた．

IC1 で特定された局所運動のように，稀にしか変動が生じない運動（レアイベント）は，ゆっくりとした変化を示すという遅い運動のイメージとは異なっている．しかし，その自己相関関数は，一般にゆっくりと緩和することが期待されるので，tICA の枠組みでは遅い時間スケールの運動とみなされることがわかる．したがって，tICA を用いると，シミュレーション中でごく稀にしか発生しないようなレアイベントを効率的に同定・抽出することができると考えられる．

3.3 IC3 によって特定された LAO 主鎖の局所運動

IC3 では残基 A15 と P16 に顕著な変位が見られ，その周辺で局所的に遅い運動が起こっていることが示唆された．図 4(a)に示した IC3 の時間発展からは，一時的な遷移が 140 ns と 470 ns

付近で二度生じているだけであることから，IC1 と同様，IC3 が表す運動もレアイベントであることがわかる．

これらの遷移において，原子レベルでどのような構造変化が起こっていたかを確認するため，当該部分の結晶構造(図 4(f))と 470 ns における構造(図 4(g))とを比較した．その結果，結晶構造では β ターンが形成された安定な構造であったものが，遷移中には複数の水素結合が切断され，局所的なアンフォールディングが 11 番目から 18 番目のアミノ酸残基に渡って起きていたことがわかる．このような局所的アンフォールディングが実際に生じていたことは，図 4(b)と(c)に示したように，2 つの主鎖二面角 D11ϕ と S18ψ が IC3 と同様の遷移挙動を示していることからも確認できる．また，該当部分で水素結合を形成していた原子間の距離 Y14O–F17N と P16O–F29N も同様の挙動を示す(図 4(d)と(e))．興味深いことに，結晶構造において P16O と水素結合を形成していた F29N は局所的なアンフォールディングの際には，その結合の相手を P16O から F17O に取り換える．そして，再度のフォールディングによってこの部分が元の構造に戻る際には，本来の相手である P16O と水素結合をきちんと形成し直していた．

以上のように，IC3 で特定された局所運動は，水素結合の切断・形成による局所的なアンフォールディング／リフォールディングであることがわかった．この部分に含まれる Y14 はリガンドと相互作用する残基であることから，観測された局所的なアンフォールディングがリガンド結合過程に関わっている可能性も考えられる．

4. おわりに

本稿では，MD シミュレーションで得られたタンパク質の複雑な運動から機能に関わる可能性が高い遅い時間スケールの運動を特定・抽出する方法として我々が提案した「時間構造に基づいた独立成分分析(tICA)」について詳しく解説した．また，タンパク質のシミュレーション結果に tICA を適用することによって，タンパク質の遅い運動を効率的に特定・抽出することができ，その動的な振る舞いを明らかにできることを示した．tICA はタンパク質のみを対象とした解析手法ではなく，実験データをはじめとした様々な時系列データにも適用可能である．今後，tICA が幅広い領域へ応用されるようになることを期待している．

参 考 文 献

甘利俊一, 村田　昇 (2002).『独立成分分析 多変量データ解析の新しい方法』, 臨時別冊・数理科学 SGC ライブラリ 18, サイエンス社, 東京.

Berendsen, H. J. C. and Hayward, S. (2000). Collective protein dynamics in relation to function, *Current Opinion in Structural Biology*, **10**, 165–169.

Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H. and Shaw, D. E. (2012). Biomolecular simulation: A computational microscope for molecular biology, *Annual Review of Biophysics*, **41**, 429–452.

渕上壮太郎 (2011). 独立成分分析 tICA によるタンパク質ダイナミクスの解析, 分子シミュレーション研究会会誌“アンサンブル”, **13**, 161–166.

Fuchigami, S., Fujisaki, H., Matsunaga, Y. and Kidera, A. (2011). Protein functional motions: Basic concepts and computational methodologies, *Advances in Chemical Physics*, **145**, 35–82.

Hayward, S. and Go, N. (1995). Collective variable description of native protein dynamics, *Annual Review of Physical Chemistry*, **46**, 223–250.

Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins, *Nature*, **450**, 964–972.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001). *Independent Component Analysis*, Wiley, New York.

ビバリネン, アーポ, カルーネン, ユハ, オヤ, エルキ (2005).『詳解 独立成分分析 信号解析の新しい世界』,

東京電機大学出版局, 東京.

Ikeguchi, M. (2004). Partial rigid-body dynamics in NPT, NPAT and NPγT ensembles for proteins and membranes, *Journal of Computational Chemistry*, **25**, 529–541.

Kitao, A. and Go, N. (1999). Investigating protein dynamics in collective coordinate space, *Current Opinion in Structural Biology*, **9**, 164–169.

Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O. and Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function, *Current Opinion in Structural Biology*, **19**, 120–127.

Lane, T. J., Shukla, D., Beauchamp, K. A. and Pande, V. S. (2013). To milliseconds and beyond: Challenges in the simulation of protein folding, *Current Opinion in Structural Biology*, **23**, 58–65.

MacKerell, A. D., Jr., Bashford, D., Bellott, M., Dunbrack, R. L., Jr., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., III, Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. and Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins, *The Journal of Physical Chemistry B*, **102**, 3586–3616.

MacKerell, A. D., Jr., Feig, M. and Brooks, C. L., III (2004). Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations, *Journal of Computational Chemistry*, **25**, 1400–1415.

Molgedey, L. and Schuster, H. G. (1994). Separation of a mixture of independence signal using time delayed correlation, *Physical Review Letters*, **72**, 3634–3637.

Naritomi, Y. and Fuchigami, S. (2011). Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions, *The Journal of Chemical Physics*, **134**, 065101.

Naritomi, Y. and Fuchigami, S. (2013). Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis, *The Journal of Chemical Physics*, **139**, 215102.

Oh, B.-H., Pandit, J., Kang, C.-H., Nikaido, K., Gokcen, S., Ames, G. F.-L. and Kim, S.-H. (1993). Three-dimensional structures of the periplasimic lysine/arginine/ornithine-binding protein with and without a ligand, *The Journal of Biological Chemistry*, **268**, 11348–11355.

Oh, B.-H., Ames, G. F.-L. and Kim, S.-H. (1994). Structural basis for multiple ligand specificity of the periplasimic lysine-, arginine-, ornithine-binding protein, *The Journal of Biological Chemistry*, **269**, 26323–26330.

Proceedings of the Institute of Statistical Mathematics Vol. 62, No. 2, 243–255 (2014) 255

Independent Component Analysis tICA to Unravel Complex Protein Motions

Sotaro Fuchigami

Graduate School of Medical Life Science, Yokohama City University

Molecular dynamics (MD) simulation is a powerful tool that is widely used to elucidate dynamic behavior of proteins and to reveal molecular mechanisms of their functions at an atomic resolution. Protein motions occur over a wide range of time scales, but not all are important for protein functions. Because time scales of the functions are generally longer, it would be reasonable to consider that slower motions of proteins are more relevant to their functions. To identify and examine such slow dynamics of proteins from simulation results, we recently proposed a method of time-structure based independent component analysis (tICA). Here, we review the approach of tICA and present the results of its application.

Key words: Protein dynamics, molecular dynamics simulation, independent component analysis, tICA, slow motion, rare event.