統計数理(2014)
第 62 巻 第 2 号 171–184

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]

相関から眺める生体分子運動の解析

桜庭 俊†

(受付 2013 年 12 月 31 日；改訂 2014 年 7 月 24 日；採択 7 月 28 日)

要　　旨

生体分子の分子シミュレーションは，大量の時系列データを出力するため，そのデータ量の削減が極めて重要になる．本稿では，分子シミュレーション分野で用いられている線形変換に基づく様々な次元削減手法と，そこに現れる相関の考え方を扱う．実例として機能モード分析，主成分分析，全相関分析，準非調和分析，独立部分空間分析などを紹介する．また，その物理学的な意味についても解説を行う．

キーワード：生体高分子，分子動力学，次元削減，独立成分分析，独立部分空間分析，エネルギー地形．

1. 生体分子のシミュレーション研究

生体高分子(タンパク質・核酸など)の分子動力学シミュレーションを行うと，原子単位での分子の運動をコンピュータ上に再現することが出来る．分子動力学シミュレーションは主に実験から得られた既知の生体分子の構造情報から出発し，各原子に掛かる力を計算し運動方程式を解くことで，原子座標の経時変化を得る．特に，各原子に掛かる力を古典力学に基づいてモデル化し計算する古典分子動力学法は，アルゴリズムと計算機の急速な進歩に支えられ，近年では生体分子の解析を支える重要な手法の一つとなっている．

分子動力学シミュレーションから得られる直接の出力は原子座標の経時変化 $\boldsymbol{x}(t)$ である(これをトラジェクトリと呼ぶ)．原子座標は原子数を N とした際に $d = 3N$ 次元，ないし並進・回転の自由度 6 を除いた $d = 3N - 6$ 次元のデータを持つ．たとえば，比較的「小さな」タンパク質分子の一つである dehydrofolate reductase は，タンパク質だけで 2400 以上の原子を持ち，その次元数は 7000 を超える．このような高次元の時系列データを人間が直接理解することは事実上不可能である．こうした理由から，分子動力学シミュレーションの結果を解釈する上では，一次データである座標情報の前処理が一番の問題となる．加えて，原子は常に熱ゆらぎの範囲でランダムな運動を続けているため，一目見て重要な運動を抽出することは困難である．したがって，ほとんどの「どうでもよい」ただの熱ゆらぎの運動を適切に除去することが，シミュレーションの結果の解析の容易さを左右することになる．この何が「どうでもよい」運動であるかの設定こそが，分子動力学シミュレーションの解析において最も本質的な問いと言ってよい．

本稿ではこの観点に基づき，高次元の情報を低次元に変換しつつできるだけ情報を残す次元削減(dimensionality reduction)の技法について扱う．次元削減に於いて何の情報を残すかの選択は，そのまま「どうでもよい」運動の選択である．次元削減の技法は統計分析や機械学習の分野で多数の文献が存在するが，本稿ではこの中で生体分子の分子シミュレーションの解析で用

†日本原子力研究開発機構 分子シミュレーション研究グループ：〒 619–0215 京都府木津川市梅美台 8–1–7

いられた技法をレビューする．生体分子シミュレーションの解析に於いては，その分子の機能とメカニズムを解明することが重要であるケースが多く，機能とかかわる様々な次元の抽出方法が考案され利用されている．

シミュレーションの解析に於いては，次元削減の結果を用いて後続の解析を行い，最終的に構造に立ち戻って生物学的な意味を考える．こうした目的で用いられる「エネルギー地形」の概念についても解説する．

2. 次元削減の方法

分子シミュレーションの解析に於いては，対応する物理的な描像，特に力学的な意味が明らかで，かつ逆変換が容易であるような変数変換に基づいた次元削減法が好まれる．線形射影は 3 節でも示すが物理的な描像がわかりやすく，また逆変換がほぼ自明に可能であることから，分子シミュレーションの次元削減に於いて最もよく利用される方法の一つである．これから挙げる次元削減法は，いずれも線形変換に基づくものである．

カルテシアン座標系で表されるシミュレーション結果の座標集団を $\boldsymbol{x}(t)$ としよう．$\boldsymbol{x}$ は d 次元の縦ベクトルである．このベクトルを $m(< d)$ 次元に縮約する線形射影を考え，射影に用いる行列を V とする．V は $m \times d$ のサイズを持つ行列であり，その列ベクトルを $\boldsymbol{v}_1^T, \boldsymbol{v}_2^T, \ldots, \boldsymbol{v}_m^T$ とする．ここで，$\cdot^T$ は行列の転置，あるいは列ベクトルと行ベクトルとの相互変換を表す．射影された後の座標 $\boldsymbol{p}(t) = \{p_1(t), p_2(t), \ldots, p_m(t)\}$ は

$$\boldsymbol{p}(t) = V\boldsymbol{x}(t) \tag{2.1}$$

として表される．以降の説明では，断りなく t を省略し表す．$\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m$ が線形独立であるとき，これらのベクトルで張られる部分空間での運動が削減された次元に於ける分子の運動に対応する．この運動，あるいは射影結果の m 次元空間上の運動が，できるだけ元来のシミュレーション結果の情報を残すように次元を削減することが求められる．

射影後の各次元は $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m$ 方向への運動を表す．$\boldsymbol{v}_i$ は d 次元のベクトルである．このベクトルは全原子座標の運動に匹敵する次元を持つため，ベクトルを各原子の運動方向として可視化することが出来る．図 1 に実際のシミュレーション結果の解析から得られる射影ベクトルの例を示す．射影ベクトルは生体分子内の複数の原子の集団的な運動を記述する．こうして射影ベクトルの向きを原子単位の向きに戻すことで，観測された運動が構造生物学的にどのような原子の運動に対応するかを説明する．

2.1 応答解析

シミュレーション対象となる生体分子について機能を果たす上での大まかな動きが明確になっている場合には，その動きと関連するような運動を知りたくなる．例えば，タンパク質のある部位の体積の変化が機能と関連することが既に明らかであるとき，それに対して応答するような運動を調べる，といった利用法である．Hub and de Groot (2009) で提唱された機能モード分析 (functional mode analysis; FMA) は，こうした応答を起こす次元を効果的に抽出する手法である．対象とする運動に対し最も強く応答する方向を求め他を無視することで，次元削減を実現する．FMA は削減先の次元が $m = 1$ となる場合のみを定義している．

応答する対象となる指標 f が，座標の関数 $f = f(\boldsymbol{x})$ として表されるとしよう．次元削減の指標として，ピアソンの相関係数

$$P(p_1, f) = \frac{\langle p_1 f\rangle - \langle p_1\rangle \langle f\rangle}{\sqrt{(\langle p_1^2\rangle - \langle p_1\rangle^2)(\langle f^2\rangle - \langle f\rangle^2)}}, \tag{2.2}$$

図 1. 線形射影のベクトルは原子の集団運動を表す．T4 lysozyme タンパク質(灰色)の古典分子動力学シミュレーション結果に PCA を適用した場合の，第 1 主成分の射影ベクトルを可視化した．分子構造が「閉じる」向きの集団運動が示されている．

を導入する．ここで，$\langle\cdot\rangle$ は値の時間平均である．ピアソンの相関係数を最大化するような射影ベクトル $\boldsymbol{v}$ が，求める方向である．

$$\boldsymbol{v}_{\mathrm{FMA}} = \arg\max_{\boldsymbol{v}_1} P(p_1, f) = \arg\max_{\boldsymbol{v}_1} P(\boldsymbol{v}_1^T \boldsymbol{x}, f). \tag{2.3}$$

この問題は，簡単な線形方程式として解くことができる．著者らはこれに加え，アンサンブル平均を考慮した拡張や，線形相関であるピアソンの相関係数ではなく非線形な相関についても扱えるよう，相互情報量を最大化する形式の FMA も定義している．

Hub and de Groot(2009)ではこの FMA を T4 lysozyme タンパク質に適用し，リガンド結合部位の空隙の体積に応答する運動を探しその運動の性質を調査している．また Trp-cage タンパク質に適用し，疎水性残基の表面積に応答する次元を探すことで，unfolding と呼ばれる立体構造が壊れる過程を集団運動として可視化することに成功している．

2.2 内在的な情報を抽出する

シミュレーションでの解析を行う対象となる生体分子は，しばしばメカニズムが全くわかっていないか，極めて限られた範囲しか知られていないケースが多い．このような生体分子を解析するには，どのような運動に着目するのが適切であろうか．

統計学でよく利用される主成分分析(principal component analysis; PCA)が，こうした内在的な運動を抽出する手法の一つとしてシミュレーション分野でも広く用いられている．PCA は分散共分散行列 $A = \langle \boldsymbol{x} - \langle \boldsymbol{x} \rangle \rangle \langle \boldsymbol{x}^T - \langle \boldsymbol{x}^T \rangle \rangle$ の対角化により定義される．

$$W \Lambda W^T = A. \tag{2.4}$$

ここで Λ は対角行列であり，A が半正定値対称行列であるため Λ の対角要素は非負実数となる．W は直交行列である．一般性を失うこと無く，$\Lambda_{ii} \geq \Lambda_{jj} (i < j)$ とするとき，この W の構成行のうち，最初の m 行がそれぞれ PCA での射影行列 V_{PCA} の列に対応する．

$$V_{\mathrm{PCA},kl} = W_{lk} = W_{kl}^T. \tag{2.5}$$

外部パラメータに最もよく応答する 1 次元の情報を抽出する FMA と異なり，PCA では複数の

次元が抽出されることに注意されたい．

PCA の物理学的な意味は，射影後の m 次元部分空間での分散を考えるとわかりやすい．PCA は射影後の分散を最大化するような方向のうち，射影ベクトルどうしが正規直交系を為すようなものを探す．

$$\mathcal{L}_{\mathrm{PCA}}(V) = -\left\langle \|\boldsymbol{p} - \langle \boldsymbol{p} \rangle\|^2 \right\rangle = -\left\langle \|V\boldsymbol{x} - \langle V\boldsymbol{x} \rangle\|^2 \right\rangle . \tag{2.6}$$

この関数は，$VV^T = I_{m\times m}$ の制約のもとでは $V = V_{\mathrm{PCA}}$ で最小値を持つ（$I_{m\times m}$ は $m \times m$ の単位行列）．したがって，PCA は構造変化に最も大きく寄与する次元を取り出し，運動のサイズが小さい運動を無視することになる．また，分散・共分散行列を対角化していることから，射影後の 2 次までの相関を取り除いている，とも表現することが出来る．

PCA の生体分子への応用は，Kitao et al.（1991），García（1992），Amadei et al.（1993）などにより導入された．PCA は，その物理的な意味から示唆されるように，タンパク質の「大まかな」運動を小数の次元で表すために最適な手法である．例えば T4 lysozyme タンパク質のシミュレーションから得られたトラジェクトリについて，de Groot et al.（1998）はタンパク質のドメインと呼ばれるタンパク質を構成するパーツ単位の運動が PCA により解析できることを報告している．このドメイン運動は T4 lysozyme の機能である加水分解との関連が深いものであった．

2.3 非ガウス性を使って次元を削減する

実際には，生体分子にはスケールの小さな運動から大きな運動まで存在し，大きな運動のみがタンパク質の機能にとって重要とは限らない．より小さな運動であっても，「特徴的」な運動を抽出する方法はあるだろうか．「特徴的」な運動の定義として，次のようなものを考える．生体分子は様々な機能を持つが，多くの場合，他の分子と相互作用を起こすことで機能を発揮する．例えば，細胞膜表面にある受容体タンパク質は，細胞外の他のタンパク質や低分子化合物（リガンドと呼ばれる）と結合し，その情報を細胞内に伝える．この時，生体分子は結合前と結合後で異なる構造を取るが，この 2 構造はリガンド分子が存在しない際にも構造ゆらぎの範囲で確率的に存在すると考えられている．したがって，単一の安定点を持たず，複数の準安定状態（meta-stable states）が見えるような運動の方向があれば，それを特徴的とするのはどうだろうか．具体的には，安定点の周りの分子運動はおおよそ調和振動子として近似できることが知られており，その分布がガウス分布となる（このことについてのより詳細な議論は 3 節で行う）．射影後の分布がガウス分布から離れた結果となるベクトルの方向を探すことで，複数の準安定状態を探すことは出来るだろうか．

これは，次のようにも捉えることが出来る．一般に，独立な多数の確率変数の平均を取ると，中心極限定理から一定の条件の下でその和はガウス分布に収束する．同様に，高次元の独立な確率変数に対し，ランダムなベクトルとの内積を取ると，それは次元が増えるに従いガウス分布に漸近していく．言い換えると，高次元で適当な射影を取ると，ほぼ確実にガウス分布様の分布になると言える（より詳しくは，例えば Hyvärinen et al., 2001 の 8 章などを参照されたい）．したがって，射影するとガウス分布になってしまうような方向は，「特徴的でない」，どうでもよい方向であり削減できる，というものである．

こうした考えに基づき，サンプル分布から複数の準安定状態が見えるような方向，あるいはガウス分布に従わない方向を見つける方法は，信号処理・機械学習の分野で射影追跡法（projection pursuit; Friedman and Tukey, 1974）や非ガウス性分析（non-Gaussian component analysis; Blanchard et al., 2006）として知られている．独立成分分析（independent component analysis; ICA, Herault and Jutten, 1986; Comon, 1994）は直接的には非ガウス成分を抽出するものではないが，非ガウス性と密接な関連を持つことが知られている（Hyvärinen et al., 2001）．分子シミュレー

ションの解析では，後述するエネルギー面を考える際に都合が良いことから，ICA に基づいたアルゴリズムが近年積極的に用いられている．ICA では，複数のベクトルを用いて射影した結果の分布が，統計的に独立である，というモデルを立てる．

$$\rho(\boldsymbol{x}) = \rho_1(p_1)\rho_2(p_2)\dots\rho_d(p_d), \tag{2.7}$$

$$\rho(\boldsymbol{x}_0) = \langle \delta(\boldsymbol{x}(t) - \boldsymbol{x}_0) \rangle , \tag{2.8}$$

$$\rho_i(p_i') = \int \delta(\boldsymbol{v}_i^T \boldsymbol{x} - p_i')\rho(\boldsymbol{x}) d\boldsymbol{x}. \tag{2.9}$$

ここで，ρ はシミュレーション結果の $\boldsymbol{x}$ 上での確率分布，ρ_i は $\boldsymbol{v}_i$ ベクトル上への射影結果の確率分布を表す．ここで $\delta(\cdot)$ はディラックのデルタ関数であり，$\int f(\boldsymbol{x})\delta(\boldsymbol{x})d\boldsymbol{x} = f(\boldsymbol{0})$ を満たす．ρ は d 次元の座標を取る関数，ρ_i は 1 次元の射影座標を取る関数となる．多次元のガウス分布は 1 次元のガウス分布の積として書けるが，逆に非ガウス的な分布はガウス分布の積だけでは決して表されないため，ガウス性を強く持つ次元が自動的に分離される．したがって，射影された結果の分布がほぼガウス分布となる分布を取り除くことで，注目する次元を絞り込むことが出来る．こうして残った次元のうちには，複数の準安定状態を持つ次元が含まれる．これが ICA に基づく分子シミュレーション結果の解析の大枠の考え方である．

Lange and Grubmüller (2008) による全相関分析 (full correlation analysis; FCA) は，ICA を分子シミュレーション分野に応用した先駆的な研究である．FCA では，射影後の確率分布のもつ相互情報量 I を最小化することで (2.7) に示された分解を行う：

$$I[p_1, p_2, p_3, \dots, p_d] = \left(\sum_{i=1}^{d} H[\rho_i] \right) - H[\rho], \tag{2.10}$$

$$H[\rho] = -\int \rho(\boldsymbol{p}) \log \rho(\boldsymbol{p}) d\boldsymbol{p}, \tag{2.11}$$

$$H[\rho_i] = -\int \rho_i(p_i) \log \rho_i(p_i) dp_i. \tag{2.12}$$

$H[\rho]$ は確率分布関数 ρ のもつ情報エントロピーを表す汎関数である．FCA を T4 lysozyme と呼ばれるタンパク質に適用することで，タンパク質運動が実際に準安定状態を持つこと，また FCA では検出が難しかった準安定状態を効率良く計算から見つけ出すことができることを示した．また，neurotensin タンパク質の運動を同様に解析し，その非調和性などを報告している．他，モードのもつ性質を PCA との差異などと合わせて報告している．

Ramanathan et al. (2011) は ICA を用いた解析として，準非調和分析 (quasi anharmonic analysis; QAA) を提唱した．Ramanathan らの研究は特に準安定状態の検出と，その準安定状態間の遷移の階層性に着目したものである．分布間の確率分布の独立性を表す指標として，4 次のキュムラントと呼ばれる量を導入する．射影結果の平均が 0, 分散が 1 になるように変数 (あるいは射影ベクトル) を正規化すると，4 次キュムラントは次のように表される．

$$\mathrm{Cumu}(p_i, p_j, p_k, p_l) = \langle p_i p_j p_k p_l \rangle - \langle p_i p_j \rangle \langle p_k p_l \rangle - \langle p_i p_k \rangle \langle p_j p_l \rangle - \langle p_i p_l \rangle \langle p_j p_k \rangle . \tag{2.13}$$

4 次キュムラントはその全ての入力変数のうち，互いに独立な変数の組が一つでも存在する時に 0 になるため，確率分布の独立性の判定法として働く．この 4 次キュムラントの二乗和を最小化して ICA を達成する方法が JADE (joint approximate diagonalization of eigenmatrices; Cardoso, 1999) アルゴリズムとして知られており，QAA はこの手法に基づく．最小化対象の関数は

$$\mathcal{L}_{\mathrm{JADE}}(V) = \sum_{ijkl \neq iikl} \mathrm{Cumu}(p_i, p_j, p_k, p_l)^2. \tag{2.14}$$

となる．全ての運動が互いに統計的に独立であるとき，関数 $\mathcal{L}_{\mathrm{JADE}}(V)$ は最小値 0 を取るため，この値の最小化により ICA が達成できる．PCA では座標の 2 次量を無相関にすることで各次元どうしの相関を消し次元削減を行ったが，JADE（QAA）では 4 次のキュムラントを無相関にすることで同様の目的を達成する．より高次の相関までを扱うことで，低次の統計量では見えなかった相関を消去している．

QAA により見出された非ガウス的な特徴を持つ運動は，ubiquitin タンパク質での頻度の低い準安定構造を発見し，また T4 lysozyme に於いて基質結合部位の制御構造を解き明かした．更に，cyclophilin A での構造遷移の中間状態を見つけ出した．さらに，著者らはタンパク質の準安定構造の間には階層性があり，大局的な準安定構造をクラスターに分けクラスター内部を調査すると，クラスター内部にも複数の準安定構造が見え，それらが QAA でうまく分離できることを指摘している．詳細は Ramanathan et al.（2011）ならびに Savol et al.（2011）を参照されたい．

「非ガウス性」を，分布の計算結果の非ガウス性ではなく，時間方向の情報を用いて決めることも出来る．ICA が元々は音声処理の文脈で発達したこともあり，時間方向の情報を用いる考えは信号処理の分野では ICA の登場直後から頻繁に用いられてきた．SOBI（second-order blind identification; Belouchrani, 1997）や TDSEP（temporal decorrelation source separation; Ziehe, 1998）は，射影結果の時間相関が消失するような方向を探す形で独立成分分析を行う．同様の「時間方向の情報を使う」という考えは生体分子のシミュレーション分野でも行われており，時間構造に基づく独立成分分析（time-structure based independent component analysis; tICA, Naritomi and Fuchigami, 2011）や緩和モード解析（relaxation mode analysis; RMA, Mitsutake et al., 2011）が提案されている．両者に共通する特徴は，特定の時間スケールを持つ運動を抽出出来るという点である．これは，異なる時間スケールを持った運動が共存する生体分子では，特に有用な手法となる．tICA はある時間遅れを設定した場合の射影後の座標の相関関数が最大となるような運動を探すことで情報を抽出する．tICA はリジン・アルギニン・オルニチン結合タンパク質に対し適用され，遅く互いに高い独立性を持った運動を効率よく抽出した．RMA は ICA と独立に提案された手法であるが，相互の時間相関が 0 であり，自己相関の緩和時間が一定の時間スケールに収まるような運動を探すことで情報を抽出する．小ペプチドの met-enkephalin に RMA を適用した結果では，安定な構造間の遷移に対応した遅い運動を抽出することができた．tICA, RMA についてのさらなる詳細は，本特集記事内にある渕上氏，高野氏の解説をそれぞれ参照されたい．

2.4 相関そのものに着目する

ICA モデルでは，射影後の確率分布は互いに独立であり，積に完全に分解可能である，と扱われていた．しかしながら，確率分布が独立であるという事は，ある次元での 2 つの準安定状態間の遷移は，他の次元での準安定構造の間の構造遷移と無相関であり影響を及ぼさない，という仮定に基づいている．生体分子は様々な機能を発揮するが，たとえば複数の分子との結合を入力として何らかの出力を与えるようなレセプター分子には，この仮定は不自然であると言える．入力による構造変化が，他の構造変化に伝わらないモデルとなっているためである．実際に，Lange and Grubmüller（2008）では FCA により検出された準安定状態の間に相関が見られるケースを報告している．

ICA モデルをわずかに変更することで，より自然なモデルを導入することは出来るだろうか．信号処理の分野で知られる独立部分空間分析（independent subspace analysis; ISA, Cardoso, 1998, 1999; Hyvärinen and Hoyer, 2000）はこのようなモデルの一つであると言える．ISA モデルに於いては，原子座標の分布関数が，互いに独立な部分空間での分布関数の積によって表される，と仮定する．例えば，対象の確率分布が 3 次元の部分空間と 2 次元の部分空間，そして

$d-5$ 個の 1 次元の部分空間の直積で表されると仮定すると，次のような式となる．

$$\rho(\boldsymbol{x}) = \rho_{123}(p_1, p_2, p_3)\rho_{45}(p_4, p_5)\rho_6(p_6)\dots\rho_d(p_d). \tag{2.15}$$

この場合，射影後の座標として p_1, p_2, p_3 で表されている，$\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ で張られる部分空間内の運動は互いに相関していても構わない．このような相関する次元の「ブロック」を導入することで，ISA は ICA を一般化する．ICA のモデルは，全次元が 1 次元の部分空間の直積で表される場合の ISA であると言える．ブロックで示される確率分布の部分空間がある，ということは，その部分空間内の運動が強い相関を持つため，単純な 1 次元どうしの運動に分解できないことを示すことになる．

Sakuraba et al.(2010)ではこの ISA に基づき，タンパク質 T4 lysozyme の運動を解析した．この研究では ISA の分解を行うため，ICA では分解不可能な部分空間を自動的に見つけ出しながら空間を分割する，SJADE(subspace JADE; Theis, 2006)アルゴリズムが利用された．SJADE アルゴリズムは 2.3 節で示した JADE アルゴリズムに，相関を持つ次元ブロックの検出を追加したものである．特に強い相関を持つ部分空間とその運動の組を解析した結果，これら相関した運動は時間遅れを伴って相関する相手の運動に応答しており，タンパク質機能と関連する運動であることを報告し，また変異体実験の結果などとも符合することを示している．

3. エネルギー地形

シミュレーションが定温・定積条件で実施されている場合，無限時間のシミュレーションを行った際の確率分布はポテンシャルエネルギーによって表されることが統計物理学から知られている．特に生体分子の場合は，ポテンシャルエネルギー面は「エネルギー地形」(energy landscape)とも呼ばれる．

$$\rho(\boldsymbol{x}) \propto \exp\left(-\frac{U(\boldsymbol{x})}{k_B T}\right). \tag{3.1}$$

ここで k_B はボルツマン定数，T は系の温度である．このことは逆に，系のポテンシャル関数を，サンプル分布から近似的に求めることができることを表している．

$$U(\boldsymbol{x}) = U_0 - k_B T \log \rho(\boldsymbol{x}). \tag{3.2}$$

U_0 は任意定数である．次元削減を行った場合には，同様に削減された後の次元でのポテンシャル関数を考えることが出来る．縮小された空間でのポテンシャル関数を考えることで，分子の様々な運動を近似的に表すことが可能である．このような，次元削減後のポテンシャル関数を，平均力ポテンシャル(potential of mean force)と呼ぶ．m 次元空間への線形射影の場合，平均力ポテンシャルは次のように定義される．

$$U(p'_1, p'_2, \dots, p'_m) = U_0 - k_B T \int \log \rho(\boldsymbol{x}) \prod_{i=1}^{m} \delta(p_i - p'_i)\, d\boldsymbol{x} \tag{3.3}$$

$$= U_0 - k_B T \int \log \rho(\boldsymbol{x}) \prod_{i=1}^{m} \delta(\boldsymbol{v}_i^T \boldsymbol{x} - p'_i)\, d\boldsymbol{x}. \tag{3.4}$$

こうした平均力ポテンシャルがどのような構造を持っているかを知ることで，生体分子が起こしうる構造変化のパターンを知ることが出来る．このような次元削減された平均力ポテンシャルの構造もまた，近似ではあるが生体分子のポテンシャル面をよく反映していることから，エネルギー地形と呼ばれている．次元削減後のエネルギー地形が元のポテンシャル面を「良く」近似している場合，生体分子の運動はより単純な地形での運動，あるいはより少数の次元におけ

る運動として扱うことができ，メカニズムの理解が容易になる．次元削減後のエネルギー地形が「良い」近似として成立するためには，元の高次元での分子の振る舞いをうまく再現あるいは説明するように，注意深く確率モデルを選ばなければならない．

3.1 主成分分析

PCA, ICA, ISA の各分析手法は，それぞれ異なるエネルギー地形をモデルとして設定して解析を実施していることに対応する．2.3 節で述べたように，PCA では射影方向の決定の際に，3 次以上のキュムラントを無視する．これは，最大で 2 次のキュムラントしか持たない分布関数で，確率分布を近似していることに相当する．このような関数は多次元ガウス分布で表すことが出来る．

$$\rho(p_1, p_2, \ldots, p_d) \approx \frac{1}{\sqrt{(2\pi)^d}\prod_{i=1}^{d}\sigma_i} \exp\left(-\frac{1}{2}\prod_{i=1}^{d}\sigma_i^{-2}(p_i-\mu_i)^2\right), \tag{3.5}$$

$$U_{\mathrm{PCA}}(p_1, p_2, \ldots, p_d) \approx U_0 + \frac{1}{2}\sum_{i=1}^{d}\sigma_i^{-2}(p_i-\mu_i)^2. \tag{3.6}$$

ここで μ_i は p_i の平均値，σ_i は標準偏差である．各次元での座標の運動は座標の 2 次関数で表されるようなポテンシャルを持つ事になる．したがって，主成分分析はエネルギー地形が調和振動子で近似出来ると仮定し，それらが無相関，すなわち独立となるような方向を探し，分散の大きいものから出力していることに対応する．ポテンシャルエネルギーが 2 次関数で近似できるような地形を特に調和的と呼ぶ．実際の生体分子のエネルギー地形には，いくつかの近似が適切ではない，調和的でない次元が含まれる．このような調和的でない次元を PCA により解析すると，その射影結果はやはり調和的ではない地形となる．Kitao et al. (1998) では PCA のモデルに基づき human lysozyme タンパク質のエネルギー地形を考察し，human lysozyme のエネルギー地形は調和振動子としての近似が完全には当てはまらず，細かな複数の安定構造が存在しその遷移が遅い少数の「多重階層モード」と呼ばれる 0.5%程度の次元，同じく調和的でなく複数の安定構造が存在する 4.5%程度の「単階層モード」，そしてほぼ調和振動子と近似できる 95%の「調和的モード」に大別できる，とした．

3.2 独立成分分析，全相関分析，準非調和分析

ICA に属する一群の手法をモデルとして用いる場合，平均力ポテンシャルは 1 次元のポテンシャルの和として表される．

$$U_{\mathrm{ICA}}(p_1, p_2, \ldots, p_d) = U_0 - k_B T\sum_{i}\log\rho_i(p_i). \tag{3.7}$$

ICA のモデルでは，特定の p_i でのエネルギー地形が非調和的であるような状況を陽に認めている．これにより，ある次元が複数の準安定状態を持つ状況をうまく記述することが出来る．特に，複数の準安定状態を持つ次元そのものが複数あり，それらの分散が同一となるような場合に，この差は顕著となる．PCA では次元の分離が困難であり，準安定状態を持つ次元の運動を纏めて考えなければならないのに対し，ICA ではそれらの次元を分離することで，エネルギー地形の理解を容易にすることができるためである．

PCA と ICA でエネルギー地形がどう表されるかの違いは，図 2(a) に示すような人工的な例で顕著に示せるであろう．分散がほぼ同一となる，2 方向にそれぞれ 2 状態を持つようなエネルギー地形を考える．この地形上でのシミュレーションの結果は図 2(b) に示すようなサンプル分布として表される．この結果に PCA を適用すると，ちょうど分布の対角線上に射影軸が取ら

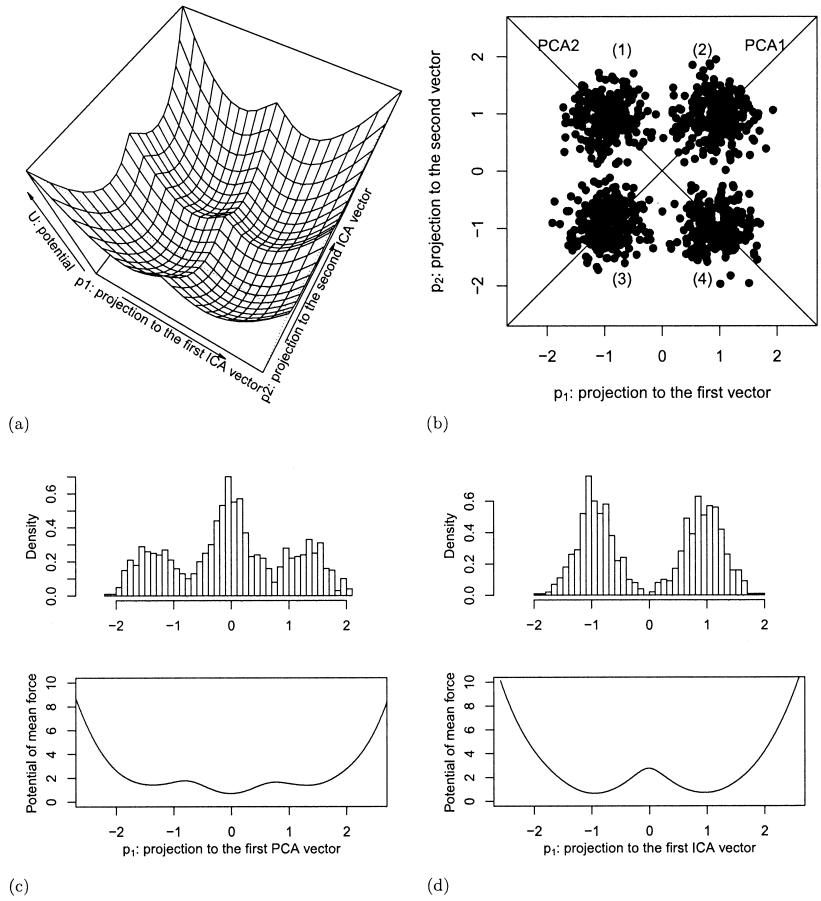

図 2. 非ガウス性を持つエネルギー地形が複数存在する場合，PCA と ICA の差は顕著となる．(a)はシミュレーション対象のエネルギー地形，(b)は対応するサンプル分布を表す(サンプル分布は平均が 0, 分散が 1 となるように正規化している)．サンプル分布は(1)–(4)までのクラスターを持つ．このサンプル分布に対し PCA を適用すると，(b)の対角線上に示した軸上に PCA の射影方向が設定される．一方，ICA を適用した場合には(b)の縦横軸方向に ICA の射影方向が設定される．PCA での第 1 ベクトルにサンプル分布を射影し，そこから平均力ポテンシャルを推定すると(c)のような分布と平均力ポテンシャルになる．PCA では 2 つの平均力ポテンシャルが共に 3 つの準安定状態を持つ．同様の操作を ICA に行うと(d)のようになり，2 つの平均力ポテンシャルはそれぞれ 2 つの準安定状態を持つ．

れる．この射影軸に沿ってサンプル分布を射影し，平均力ポテンシャルを求めると図 2(c)のようなポテンシャル面が得られる．2 つの次元がそれぞれ左右と中央にわずかに他より安定な状態を持つエネルギー地形である．一方，ICA を適用すると図 2(d)のような射影後の確率分布が得られる．ICA では 2 方向にそれぞれ 2 状態が存在するエネルギー地形の構造を再現している．

このエネルギー地形上での，構造変化を考える．図 2(b)で隣り合った準安定状態間の遷移，たとえば(1)から(2)への構造遷移は活性化障壁が低いため起こりやすく，離れた準安定状態間

の遷移，たとえば(2)から(3)への構造遷移は起こりにくい．一方，これが PCA に射影されると，(3)は図 2(c)で左端，(1), (4)は中央，(2)は右端に対応する．射影後のエネルギー地形の上では左右はなだらかに接続されているが，左右端から反対側の端への遷移は地形から想像するよりもずっと起こりにくくなる．このような運動はまさに PCA で「多重階層」と呼ばれていたものを再現している．PCA で見えていた「多重階層モード」は，ICA モデルに於いては複数の次元での運動に分解される可能性がある（他の議論は Lange and Grubmüller, 2008 や Ramanathan et al., 2011 などを参考にされたい）．今後，ICA モデルでのタンパク質のエネルギー地形の解析が増えることで，「多重階層モード」の正体が徐々に明らかになるのではないかと考えられる．

現行の確率分布を用いた ICA のアルゴリズムは，現時点では計算量ならびに数値的安定性の都合から，FCA に於いても QAA に於いても PCA により分散の小さな次元を削減した後に手法が適用されている．これは先述の PCA でのエネルギー地形の解析から，下位のほとんどの次元が調和的であることにより正当化されている．しかしながら，2.3 節で示したように，高次元の空間でのランダム射影は高確率で調和的な次元を与える．したがって，PCA により次元を削減する操作なしに ICA を適用することで，現在では調和的な次元であると考えられている空間からより多くの情報を抽出できる可能性がある．ICA アルゴリズムの生体分子シミュレーションへの今後の応用では，高次元の構造空間に対して適用可能なアルゴリズムの開発と，非調和的な次元のうちで実際に重要な次元をさらに絞り込む操作などが重要となるだろう．

ICA のモデルでは p_i 上の運動は $p_j (i \neq j)$ 上の運動と関連を持たない．したがって，i 番目の次元で複数の準安定状態が存在しても，$j (i \neq j)$ 番目の次元での準安定状態とは相互作用をしない．非ガウス的な分布を示す次元が多数存在した場合も，他の基準で興味を持たない次元を選択できれば，その次元を無視し残った次元だけのエネルギー地形を考えても構わない．エネルギー地形の解析手段として見た場合，常に最大 1 次元のエネルギー地形を議論すれば良いという点で，これは大きな利点である．その一方で，確率分布が完全な 1 次元分布の積で書けないようなケースでは，モデルを適合させるための近似が必要となる．

3.3 独立部分空間分析

ISA をモデルとする場合，平均力ポテンシャルはポテンシャルの「次元ブロック」を持つことが許可される．例えば式(2.15)で示したような 3×1 次元，2×1 次元および $1 \times (d-5)$ 次元に分割されるケースでは，

$$U_{\mathrm{ISA}}(p_1, p_2, \ldots, p_d) = U_0 - k_B T \log \rho_{123}(p_1, p_2, p_3) - k_B T \log \rho_{45}(p_4, p_5) - k_B T \log \rho_6(p_6) - \ldots - k_B T \log \rho_d(p_d), \tag{3.8}$$

のように構成される．ポテンシャルは次元ブロックに分割され，次元ブロックをまたぐポテンシャルは互いに相関を持たない．これにより，1 次元では表現できない，例えば一直線上に並ばない多数の準安定状態の間の遷移などを表すことが可能となる．

ICA と ISA の差が顕著となる人工的な例を図 3(a)に示す．この地形から得られたサンプル分布に ICA を適用すると，図 3(c)のような 2 つの平均力ポテンシャルの和に分解され，その積から推定されるエネルギー地形は図 3(d)のようになる．タンパク質の構造が(1)で示される位置にある場合，(3)の構造になるためには，(2)を経由することで途中でのエネルギー障壁を低く抑えられる．また，一旦(2)に到達しなければ，(1)から(3)までの構造変化は起こせない．(1)から(2)へ遷移の構造変化は，(2)から(3)への遷移の構造変化と相関していると言える．しかしながら，こうした相関の情報は，ICA のモデルでは表現できない．ICA では各次元が完全に独立であるような分解以外が許容されないためである．一方，表現力を増やした代償として，ISA では相関を持つ次元ブロックを切り離すことができなくなる．また，実際のサンプル分布

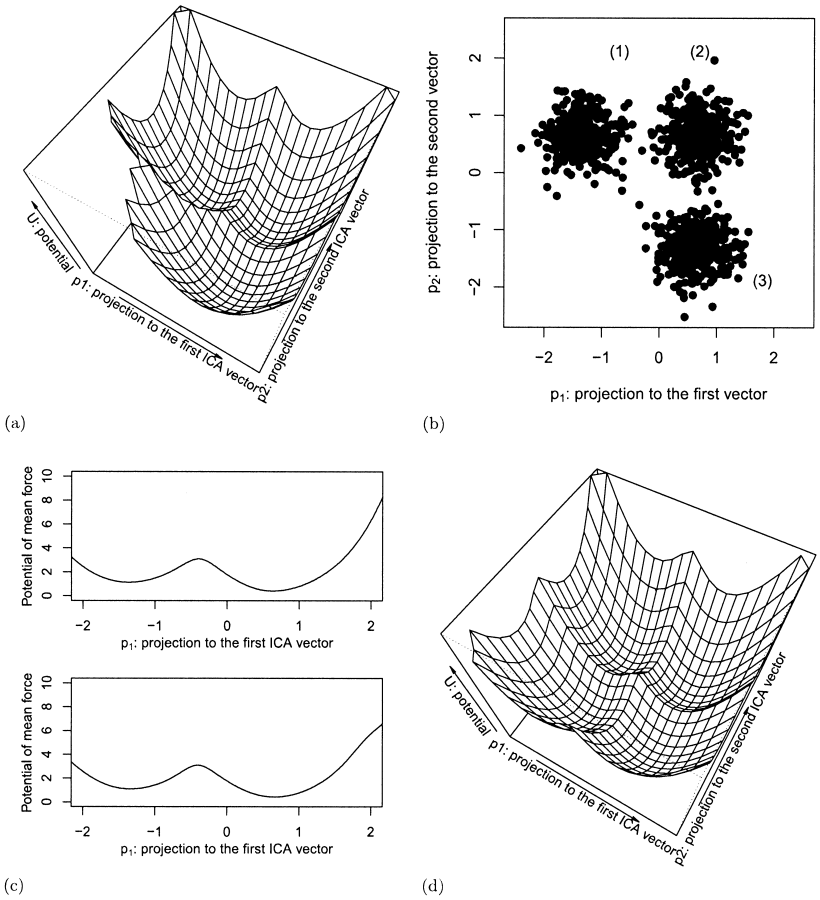

図 3. 対象とするサンプル分布のエネルギー地形が，各次元での単純な和に分解できない場合，ICA と ISA の差が顕著となる．(a)はシミュレーション対象のエネルギー地形，(b)は対応するサンプル分布を表す(サンプル分布は平均が 0，分散が 1 となるように正規化している)．サンプル分布は(1)−(3)までのクラスターを持つ．このサンプル分布に対し ICA(JADE)を適用すると，2 つの次元は(c)のような平均力ポテンシャルになる．ここから ICA のモデルで想定されているエネルギー地形を再構成すると(d)のようになり，(a)に存在しなかったエネルギー最小点が 1 つ増えてしまう．

から得られた分布は完全な分布関数では無いので，統計誤差の範囲で無相関であるものが相関を持つと判定されたり，あるいは逆のケースが起こりうる．どこまでの相関を適切な相関と認めるかについて，閾値の任意性があり，これはそのまま次元ブロックの切り出し粒度についての任意性となる．このような任意性をどのように埋めていくべきかは今後の課題となるだろう．

ISA は変数の独立性と依存性のバランスを取ったモデルである．変数同士が強固に結合したモデルは，一つの変数の値の変動が多数の変数に影響し，その挙動を説明することが難しくなる．一方，生体分子の解析では，分子内外の変化に対する何らかの相関を表現する必要がある．たとえば，ISA では，分子内運動の依存性を陽に表現したモデルを作る必要があった．新たな

次元削減の方法，エネルギー地形の近似モデルなどを作る上でも，このバランスをうまく保つ必要があるだろう．変数の独立性は，最終的な結果を人間が理解するという前提から要求され，変数の相関は解析結果と機能との関連を記述するために要求されることになる．

4. おわりに

原子単位の運動を「見る」ことが出来る分子シミュレーションの技術は近年劇的な進展を見せた．分子シミュレーションは，今後の生体分子の機能解析に於いても大きな役割を果たしていくことが期待される．分子シミュレーション技術そのものは円熟期を迎えたと言って過言ではないだろうが，大規模・複雑系の分子シミュレーション結果の解析技術はまだ発展途上である．統計・応用数理分野との分野交流によって表現力の豊富なモデルが導入され，解析技術が更に発展していくことを期待したい．その一方で，新たなモデルや手法を分子シミュレーションに導入する際には，ただ既知の手法を移植するのではなくその解析の生物学的，物理学的な意味が明瞭となるよう，注意深く定義やアルゴリズムを検討せねばならない．分子シミュレーション分野は他のシミュレーションの例に漏れず実験との協調により成り立つ分野であり，結果の生物学的，物理学的な解釈なくして協調は成り立たないためである．

参 考 文 献

Amadei, A., Linssen, A. B. M. and Berendsen, H. J. C. (1993). Essential dynamics of proteins, *Proteins*, **17**, 412–425.

Belouchrani, A., Abed-Meraim, K. and Cardoso, J.-F. (1997). A blind source separation technique using second-order statistics, *IEEE Transactions on Signal Processing*, **45**, 434–444.

Blanchard, G., Kawanabe, M., Sugiyama, M., Spokoiny, V. and Müller, K.-R. (2006). In search of non-Gaussian components of a high-dimensional distribution, *Journal of Machine Learning Research*, **7**, 247–282.

Cardoso, J. F. (1998). Multidimensional independent component analysis, *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, 1941–1944, Seattle, USA.

Cardoso, J. F. (1999). High-order contrasts for independent component analysis, *Neural Computation*, **11**, 157–192.

Comon, P. (1994). Independent component analysis, a new concept?, *Signal Processing*, **36**, 287–314.

de Groot, B. L., Hayward, S., van Aalten, D. M. F., Amadei, A. and Berendsen, H. J. C. (1998). Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data, *Proteins*, **31**, 116–127.

Friedman, J. H. and Tukey, J. W. (1974). A Projection pursuit algorithm for exploratory data analysis, *IEEE Transactions on Computers*, **C-23**, 881–890.

García, A. E. (1992). Large-amplitude nonlinear motions in proteins, *Physical Review Letters*, **68**, 2696–2699.

Herault, J. and Jutten, C. (1986). Space or time adaptive signal processing by neural network models, *Proceedings of Neural Networks for Computing*, Vol. 151, 206–211, American Institute of Physics Conference Proceedings, Snowbird, USA.

Hub, J. S. and de Groot, B. L. (2009). Detection of functional modes in protein dynamics, *PLoS Computational Biology*, **5**, e1000480.

Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase- and shift-Invariant features by decomposition of natural images into independent feature subspaces, *Neural Computing*, **12**, 1705–1720.

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

Kitao, A., Hirata, F. and Go, N. (1991). The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum, *Chemical Physics*, **158**, 447–472.

Kitao, A., Hayward, S. and Go, N. (1998). Energy landscape of a native protein: Jumping-among-minima model, *Proteins*, **33**, 496–517.

Lange, O. F. and Grubmüller, H. (2008). Full correlation analysis of conformational protein dynamics, *Proteins*, **70**, 1294–1312.

Mitsutake, A., Iijima, H. and Takano, H. (2011). Relaxation mode analysis of a peptide system: Comparison with principal component analysis, *The Journal of Chemical Physics*, **135**, 164102.

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.

Ramanathan, A., Savol, A. J., Christopher, J., Agarwal, P. K. and Chennubhotla, C. S. (2011). Discovering conformational sub-states relevant to protein function, *PLoS ONE*, **6**, e15827.

Sakuraba, S., Joti, Y. and Kitao, A. (2010). Detecting coupled collective motions in protein by independent subspace analysis, *The Journal of Chemical Physics*, **133**, 185102.

Savol, A. J., Burger, V. M., Agarwal, P. K., Ramanathan, A. and Chennubhotla, C. S. (2011). QAARM: Quasi-anharmonic autoregressive model reveals molecular recognition pathways in ubiquitin, *Bioinformatics*, **27**, i52–i60.

Theis, F. J. (2006). Towards a general independent subspace analysis, *Advances in Neural Information Processing Systems*, Vol. 19, 1361–1368, MIT Press, Cambridge, USA.

Ziehe, A. and Müller, K. R. (1998). TDSEP – An efficient algorithm for blind separation using time structure, *Proceedings of the 8th International Conference on Artificial Neural Networks*, 675–680, Skövde, Sweden.

184 Proceedings of the Institute of Statistical Mathematics Vol. 62, No. 2, 171–184 (2014)

Dimensionality Reduction and Correlation of Biomolecular Motions

Shun Sakuraba

Molecular Modeling and Simulation Group, Japan Atomic Energy Agency

In this review, we explain dimensionality reduction techniques applied in the field of biomolecular simulation. Methods based on the linear transformation are reviewed. We also investigate how these methods use correlation for reducing the dimension of biomolecules. Our review includes functional mode analysis, principal component analysis, full correlation analysis, quasi anharmonic analysis, and independent subspace analysis. We also discuss the physical interpretation and implications of the methods.

Key words: Biopolymers, molecular dynamics, dimensionality reduction, independent component analysis, independent subspace analysis, energy landscape.