

生体分子シミュレーションの摂動解析

小山 洋平[†]

(受付 2013 年 12 月 20 日; 改訂 2014 年 7 月 18 日; 採択 7 月 22 日)

要 旨

タンパク質の機能を原子レベルから詳細に理解するためにはタンパク質主鎖や側鎖の異なる構造状態の同定, 化合物・DNA・他のタンパク質との様々な結合状態の同定, これらに対する水やイオンなどの溶媒の影響, について総合的に評価する必要がある. このような複雑な構造状態とその相互作用を明らかにするために, 原子間相互作用に対する摂動解析を用いた我々の取り組みについて解説する. まず, タンパク質内あるいは複合体内部相互作用に対する摂動解析である PEPCA (potential energy principal component analysis) を導入する. PEPCA は相互作用エネルギーに対する主成分分析を実行することで, その主成分得点によりタンパク質内あるいは複合体内部の相互作用パターンが異なる構造状態が分離され, 固有ベクトル成分から構造状態間の差異を生み出している相互作用の情報が得られる. 次に, タンパク質または複合体と溶媒間相互作用の摂動解析である DIPA (distance-dependent intermolecular perturbation analysis) を導入する. DIPA は原子間力と平均溶媒原子数の積に対して関数主成分分析を実行することで, その主成分得点により溶媒との相互作用パターンが異なる構造状態が分離され, 固有関数成分により構造状態間の相違を生み出している溶媒との距離依存的な相互作用の情報が得られる. 最後に, 10 個のアミノ酸からなるシニョリンのフォールディングシミュレーションに対して PEPCA と DIPA を適用し, 固有ベクトル成分または固有関数成分と主成分得点の散布図を同時に表示する可視化法であるバイプロットを用いた結果の解釈について解説する.

キーワード: 分子動力学シミュレーション, 摂動解析, 主成分分析, 関数主成分分析, バイプロット.

1. はじめに

生体中には様々な種類の分子が存在しているが, タンパク質が触媒反応などの機能を発現するためには特定の分子を認識し, その機能を選択的に発現する必要があると考えられる (特異性). このことからタンパク質に様々な刺激を加え, 高い応答性を示す刺激を調べることでタンパク質の機能発現メカニズムや特異性について明らかにできる可能性がある. Ming らはタンパク質構造と化合物構造の結合前後での同時分布 (joint distribution) を基準振動解析 (normal mode analysis) で解析的に求めた. 化合物の結合に伴うタンパク質構造の応答性を定量化するために, 結合前後での同時分布変化の大きさをカルバック・ライブラーダイバージェンス (Kullback-Leibler divergence; 相対エントロピー; Kullback, 1959) を用いて測った. 化合物をタンパク質表面の様々な部位に発生させ, 各部位でカルバック・ライブラーダイバージェンスを計算すると, 実際の化合物の結合部位で大きな分布変化が生じることを見出した (Ming and Wall, 2005). その後,

[†]理化学研究所 生命システム研究センター: 〒 650-0047 兵庫県神戸市中央区港島南町 1-6-5

同時分布ではなく化合物発生前後でのタンパク質の構造分布(周辺分布)を基準振動解析により求め、タンパク質の構造分布変化の大きさをカルバック・ライブラーダイバージェンスにより定量化した。様々なタンパク質と化合物のデータセットを用いて、実際の化合物の結合部位ではタンパク質構造の分布変化が大きくなり、化合物結合部位の予測に有用であることを示した(Ming and Wall, 2006)。

通常の機械とは異なり、タンパク質のような生体分子機械は自発的な動きや周りの分子との相互作用を通じて常にゆらいている。分子動力学(MD)シミュレーションを用いてこのような生体分子や溶媒のゆらぎを詳しく調べることが可能である。平衡統計力学では平衡状態に摂動が加わった新たな平衡状態でのゆらぎを、摂動前の平衡状態に摂動法を適用することで推測することが可能である。特に、摂動が“微小”であり、摂動ハミルトニアンが時間にあらわに依存しないときには、摂動前後での物理量の平均値の差が未摂動平衡状態での物理量と摂動ハミルトニアンとの共分散で表されるという静的な線形応答理論(static linear response theory)が成立する(Zwanzig, 2001)。池口らは化合物の結合を原子に一定力を加えるという摂動でモデル化し、化合物が非結合状態でのMDシミュレーションに対して静的な線形応答理論を適用することで、この摂動による平均構造変化の予測を行った(Ikeguchi et al., 2005)。この結果、実際の化合物結合前後での構造変化と良く一致すること、力の向きを多少変化させても構造変化の方向はあまり変わらないという頑強性があることを示した。

タンパク質などの生体分子の機能はその構造形成(フォールディング)や(部分的な)アンフォールディング、化合物・DNA・その他のタンパク質との様々な部位での結合・解離、それに前後する構造変化などが周囲の水やイオンなどの溶媒の影響を受けながら生じる。従って、生体分子の機能を原子レベルから詳細に理解するためにはこのような複雑な構造状態の変化を体系的に取り扱う必要がある。生体分子構造の複雑性は生体分子内および生体分子・溶媒分子間におけるファンデルワールス相互作用や静電相互作用などの原子間相互作用により生じており、複数の構造状態が存在する場合には、これらのパターンが異なっていると考えられる。このため、本稿ではどのような構造状態が存在するかを原子間相互作用により特徴付け、各状態間の差異を生み出している影響の大きな相互作用を明らかにし、それらの各状態への寄与を明らかにすることを旨とする。このためには、相互作用の生体分子構造状態への影響を総合的に評価する必要がある。生体分子構造の静的なゆらぎの情報は生体分子の構造分布から得ることができるため、相互作用の強弱を変化させたときに生体分子の構造分布がどれほど変化するかを調べることでその影響を評価することにする。構造状態間の差異に関係していない相互作用の強弱を変化させても生体分子の構造分布はあまり大きく変化しないと考えられる一方で、差異を生み出している相互作用の強弱を変化させたときには大きな構造分布変化が生じると予想される。従って、相互作用の強弱を変化させ、大きな構造分布変化を生み出す相互作用の組み合わせを探索することで、生体分子の構造状態に大きな影響を及ぼす相互作用を見出すことができると考えられる。また、これらの相互作用の各状態への影響は、相互作用の強弱を変化させたときの構造状態の比率変化により特徴付けられると予想される。

相互作用の強弱を変化させることは摂動ポテンシャルエネルギーを導入することで実現でき、このような摂動により生体分子の構造状態を理解しようとする上記のアプローチを本稿では相互作用の摂動解析と呼ぶことにする。摂動に伴う生体分子構造の分布変化はカルバック・ライブラーダイバージェンスを用いて定量化を行う。フォールディング、部分的なアンフォールディング、複合体の形成・解離、といったような基準振動解析が適用できないような大きな構造変化も扱えるように、平衡状態のMDシミュレーションの結果に対して摂動法を適用することでカルバック・ライブラーダイバージェンスを推定する。生体分子内相互作用の摂動解析(Koyama et al., 2008)と生体分子・溶媒分子間相互作用の摂動解析(Koyama et al., 2011)について紹介を

行った後、これらの摂動解析の結果がバイプロットを用いて理解できることを示す。

2. 摂動解析

2.1 摂動による生体分子構造の分布変化

生体分子の座標を q 、溶媒分子の座標を q' 、系のポテンシャルエネルギーを $V(q, q')$ とする。ここで、生体分子は複数種類あってもよい。原子数・体積・温度一定 (NVT) 条件での平衡状態における分布は逆温度を β とするカノニカル分布 $\rho(q, q') = e^{-\beta V(q, q')}/Z$ で表される。 Z は規格化定数に相当する分配関数である。摂動ポテンシャルエネルギー $\Delta V(q, q')$ を用いて系に摂動を与えることを考える。摂動後の系のポテンシャルエネルギーは $V'(q, q') = V(q, q') + \Delta V(q, q')$ となり、摂動後の分配関数を Z' とすると、摂動後の分布は

$$(2.1) \quad \rho'(q, q') = \frac{1}{Z'} e^{-\beta V'(q, q')} = \frac{e^{-\beta \Delta V(q, q')}}{\langle e^{-\beta \Delta V} \rangle} \rho(q, q')$$

となる。 $\langle A \rangle$ は未摂動平衡状態での物理量 $A(q, q')$ の平均を表す。このとき、摂動後の生体分子の構造分布は

$$(2.2) \quad \rho'(q) = \int \rho'(q, q') dq' = \frac{\langle e^{-\beta \Delta V} | q \rangle}{\langle e^{-\beta \Delta V} \rangle} \rho(q)$$

と書ける。ここで、 $\rho(q)$ は摂動前の生体分子の構造分布、 $\langle A | q \rangle$ は生体分子構造 q を固定した状態での物理量 $A(q, q')$ の未摂動平衡状態での条件付期待値を表す。摂動による分布変化をカルバック・ライブラーダイバージェンス $D(\rho'(q) || \rho(q)) \equiv \int \rho'(q) \ln \rho'(q) / \rho(q) dq$ で定量化し、式 (2.2) を代入すると

$$(2.3) \quad D(\rho'(q) || \rho(q)) = \frac{1}{2} \text{var}(\beta \langle \Delta V | q \rangle) + \dots$$

と展開できることが分かる。 $\text{var}(A)$ は未摂動平衡状態における物理量 $A(q, q')$ の分散を表す。以上のようにして、摂動ポテンシャルエネルギー ΔV による生体分子構造の分布変化は ΔV の条件付期待値の分散で近似できることが分かる。

ΔV が生体分子構造 q だけに依存する場合には $\langle \Delta V | q \rangle = \Delta V(q)$ となり、条件付期待値を計算する必要がなくなる。生体分子内の相互作用に摂動を加える場合がこれに該当する。一方で、生体分子と溶媒分子の間の相互作用に摂動を加える場合には ΔV が q と q' に依存するため、条件付期待値を具体的に評価する必要が生じる。このように、生体分子内相互作用への摂動と生体分子・溶媒分子間相互作用への摂動では、式 (2.3) の条件付期待値に対する扱いが変わるため、以下ではこの二つを分けて考える。

2.2 生体分子内相互作用の摂動解析 (PEPCA)

生体分子構造の複雑性はファンデルワールス相互作用や静電相互作用などの原子間相互作用の組み合わせにより生じている。生体分子構造への影響が大きな生体分子内相互作用を明らかにするために、その強弱を変化させる摂動解析を考える。ある生体分子内相互作用のポテンシャルエネルギーを $V_k(q)$ とするとき、摂動パラメーター λ_k を導入してエネルギー項を $(1 + \lambda_k) V_k(q)$ とすれば、 $\lambda_k > 0$ で相互作用を強め、 $\lambda_k < 0$ で弱めることができる (図 1(a))。これより、相互作用の強弱を変化させるには $\lambda_k V_k(q)$ という摂動項を加えれば良いことが分かる。生体分子内の M 個の相互作用 $\mathbf{V}(q) = (V_1(q), \dots, V_M(q))$ の強弱を変える摂動は M 個の摂動パラメーター $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_M)^T$ を用いて摂動ポテンシャルエネルギーを

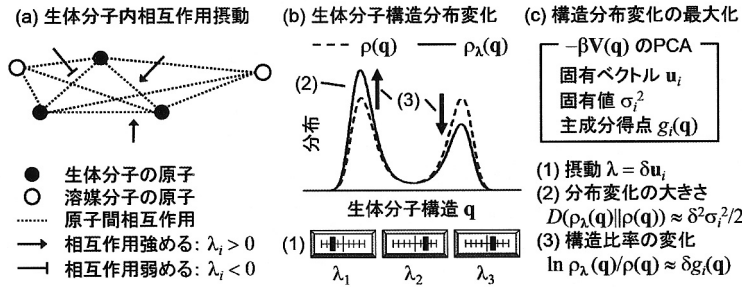


図 1. 生体分子内相互作用の摂動解析 (PEPCA).

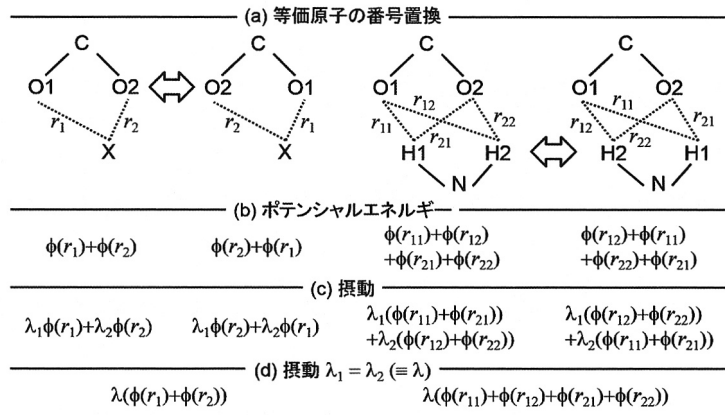


図 2. 等価原子の番号置換に対するポテンシャルエネルギーの不変性と摂動による影響.

$$(2.4) \quad \Delta V(\mathbf{q}) = \sum_{i=1}^M \lambda_i V_i(\mathbf{q})$$

とすることで実現できる.

アスパラギン酸の側鎖に含まれるカルボキシル基 CO_2 の酸素原子やアルギニンのグアニジノ基に含まれる NH_2 の水素原子のように, 等価な原子の番号の入れ替え (図 2(a)) に対して, ポテンシャルエネルギーが不変であるという置換対称性が存在する (図 2(b); Malolepsza et al., 2010). このような等価な原子との相互作用に対して異なる摂動パラメータを用いて摂動を加えると, 摂動後のポテンシャルエネルギーの置換対称性が失われてしまう (図 2(c)). 一方で, 等価な原子との相互作用のどちらか一方が生体分子の構造分布により大きな影響を及ぼすと決めることはできないため, これらの相互作用に対する摂動パラメータには同一のパラメータを用いることにする. これは置換対称性のあるポテンシャルエネルギーのグループをまとめて V_k と再定義することに相当し, このとき, 摂動後もポテンシャルエネルギーの置換対称性が保たれることが分かる (図 2(d)).

摂動の強さが一定 ($|\lambda| = \delta, \delta > 0$) という条件で生体分子内相互作用の強弱を様々に変化させたときに, 大きな構造分布変化を示す相互作用の摂動を調べることで, 生体分子構造に対して

大きな影響を及ぼす生体分子内相互作用が明らかにできると考えられる(図 1(b)). 生体分子内相互作用の摂動(2.4)による生体分子構造の分布変化の大きさ(2.3)は

$$(2.5) \quad D(\rho'(\mathbf{q})||\rho(\mathbf{q})) = \frac{1}{2} \text{var} \left(\sum_{i=1}^M \lambda_i \beta V_i(\mathbf{q}) \right) + \dots$$

$$(2.6) \quad = \frac{1}{2} \boldsymbol{\lambda}^T \text{cov}(\beta \mathbf{V}) \boldsymbol{\lambda} + \dots$$

と書ける. ここで, $\text{cov}(\beta \mathbf{V})$ は未摂動平衡状態での $\beta \mathbf{V}(\mathbf{q})$ の共分散行列を表す. $|\boldsymbol{\lambda}| = \delta$ という条件で構造分布変化(2.5)を最も大きく変化させるような $\boldsymbol{\lambda}$ を探すためには, δ が小さく, 式(2.5)の近似が成り立つという条件において式(2.5)の分散を最大化すればよいことが分かる. 式(2.5)の分散内を係数 λ_i と統計量 $\beta V_i(\mathbf{q})$ の線形結合とみなすと, $|\boldsymbol{\lambda}| = \delta$ という条件での分散の最大化は統計量 $\beta \mathbf{V}(\mathbf{q})$ または $-\beta \mathbf{V}(\mathbf{q})$ の主成分分析(principal component analysis; PCA; Jolliffe, 2002)により実行できる. 従って, 共分散行列 $\text{cov}(\beta \mathbf{V})$ の固有値を降順に σ_i^2 , 対応する固有ベクトルを \mathbf{u}_i とすると, $\boldsymbol{\lambda} = \delta \mathbf{u}_1$ のときに $|\boldsymbol{\lambda}| = \delta$ という条件で分散が最大化されることが分かる. 同様に, $\boldsymbol{\lambda} = \delta \mathbf{u}_i$ は $|\boldsymbol{\lambda}| = \delta$ および $\mathbf{u}_1^T \boldsymbol{\lambda} = 0, \dots, \mathbf{u}_{i-1}^T \boldsymbol{\lambda} = 0$ という条件での分散の最大化に対応する.

摂動 $\boldsymbol{\lambda}$ での生体分子構造の分布を $\rho_{\boldsymbol{\lambda}}(\mathbf{q})$ とする. 摂動 $\boldsymbol{\lambda} = \delta \mathbf{u}_i$ による生体分子構造の分布変化の大きさ(2.6)は

$$(2.7) \quad D(\rho_{\delta \mathbf{u}_i}(\mathbf{q})||\rho(\mathbf{q})) \approx \frac{1}{2} \delta^2 \sigma_i^2$$

となり, 共分散行列の固有値が分布変化の大きさに対応することが分かる. これより, 大きい固有値に対応する固有ベクトル(上位の固有ベクトル)により生体分子構造に大きな影響を及ぼす生体分子内相互作用が得られると考えられる. また, $g_i(\mathbf{q}) \equiv -\beta \mathbf{u}_i^T (\mathbf{V}(\mathbf{q}) - \langle \mathbf{V} \rangle)$ と定義すると, 構造 \mathbf{q} の比率変化は式(2.2), (2.4)を用いて

$$(2.8) \quad \ln \frac{\rho_{\delta \mathbf{u}_i}(\mathbf{q})}{\rho(\mathbf{q})} = \delta g_i(\mathbf{q}) + \dots$$

と書ける. ここで導入した $g_i(\mathbf{q})$ はその定義から $-\beta \mathbf{V}(\mathbf{q})$ を用いた PCA の第 i 主成分得点であることが分かる. 式(2.8)により主成分得点 $g_i(\mathbf{q})$ の符号で摂動による構造比率の増減を近似的に推定でき, 対応する固有ベクトルの相互作用がどのような構造状態に対して影響を及ぼしているかを推測することができる. この影響を可視化する手法として次節でバイプロットを用いて解説する. 以上から, ポテンシャルエネルギー $-\beta \mathbf{V}(\mathbf{q})$ の PCA (potential energy PCA; PEPCA; Koyama et al., 2008; Koyama et al., 2011) を行うことで生体分子内相互作用の摂動解析が実行できることが分かる(図 1(c)).

2.3 バイプロット

ここでは固有ベクトル成分の散布図と主成分得点の散布図を同時に表示する可視化法であるバイプロット(biplot; Gabriel, 1971; Gower and Hand, 1996)を用いることで PEPCA 固有ベクトルやその線形結合による摂動が引き起こす構造比率の変化を調べることが可能であることについて説明する. 以下では 2 次元のバイプロットで説明を行うが 3 次元のバイプロットでも同様の解釈ができる. PEPCA 固有ベクトル \mathbf{u}_i の k 番目の成分を U_{ki} とする. 第 i, j 固有ベクトル成分の散布図(U_{ki}, U_{kj})と第 i, j 主成分得点の散布図($g_i(\mathbf{q}), g_j(\mathbf{q})$)を同時に表示した 2 次元バイプロットを考える(図 3).

最初に第 i 固有ベクトルの摂動による影響について考える(図 3(a)). 摂動 $\boldsymbol{\lambda} = \delta \mathbf{u}_i$ により k 番目の相互作用のポテンシャルエネルギーは $(1 + \delta U_{ki}) V_k(\mathbf{q})$ となる. $U_{ki} > 0$ であれば相互作

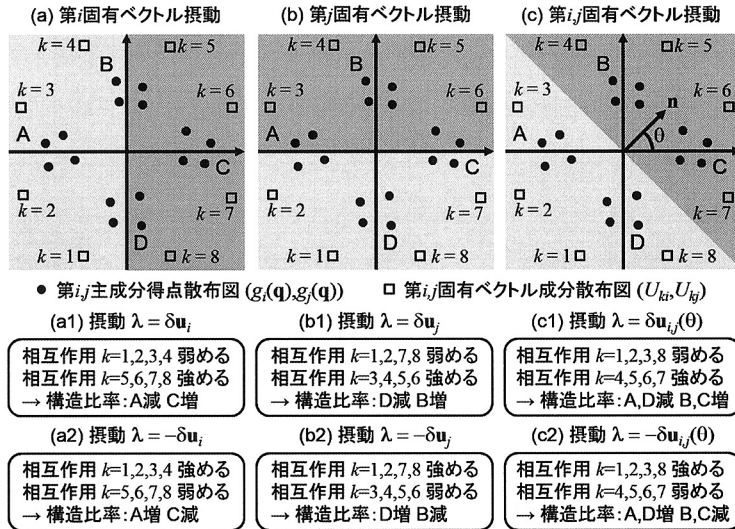


図 3. バイプロット. 摂動による構造比率の増加・減少は式(2.8)の近似に基づく.

用が強まり, $U_{ki} < 0$ であれば弱まる. また, 式(2.8)の近似の下で $g_i(\mathbf{q}) > 0$ であれば構造比率が増加し, $g_i(\mathbf{q}) < 0$ であれば減少する. このようにして, バイプロット上で左側の相互作用を弱め, 右側の相互作用を強めると左側の主成分得点に対応する構造の比率が減少し, 右側の構造の比率が増加することが分かる(図 3(a1)). 同様にして, 摂動 $\lambda = -\delta u_i$ を考えることで左側の相互作用を強め, 右側の相互作用を弱めると, 左側の構造の比率が増加し, 右側の構造の比率が減少することが分かる(図 3(a2)). 第 j 固有ベクトルの摂動による構造比率変化についても同様に, 上記の左右を下上に置き換えればよい(図 3(b)).

また, 固有ベクトルの線形結合 $\mathbf{u}_{i,j}(\theta) \equiv \cos \theta \mathbf{u}_i + \sin \theta \mathbf{u}_j, -\pi \leq \theta < \pi$ による摂動 $\lambda = \delta u_{i,j}(\theta)$ を導入することで左右や下上で分離される領域の構造比率変化だけではなく, 原点を通り単位法線ベクトル $\mathbf{n} \equiv (\cos \theta, \sin \theta)$ を持つ直線により分離される領域についても同様の考え方ができることが次のようにして分かる(図 3(c)). $\mathbf{u}_{i,j}(\theta)$ の k 番目の成分は $\mathbf{n} \cdot (U_{ki}, U_{kj})$ と書け, 固有ベクトル成分 (U_{ki}, U_{kj}) を軸 \mathbf{n} へ射影することで得られる. また, 摂動 $\lambda = \delta u_{i,j}(\theta)$ による構造比率変化は式(2.2), (2.4)を用いると $g(\mathbf{q}) \equiv \cos \theta g_i(\mathbf{q}) + \sin \theta g_j(\mathbf{q})$ と定義することにより式(2.8)と同様の式が成り立つことが分かる. $g(\mathbf{q}) = \mathbf{n} \cdot (g_i(\mathbf{q}), g_j(\mathbf{q}))$ と書けることから $g(\mathbf{q})$ は主成分得点 ($g_i(\mathbf{q}), g_j(\mathbf{q})$) を \mathbf{n} に射影することで得られる. 以上から, 固有ベクトル成分 (U_{ki}, U_{kj}) を軸 \mathbf{n} に射影した値が正である相互作用を強め, 負である相互作用を弱めると主成分得点 ($g_i(\mathbf{q}), g_j(\mathbf{q})$) を軸 \mathbf{n} に射影した値が正である構造の比率が増加し, 負である構造の比率が減少することが分かる(図 3(c1)).

以上の結果を利用することにより, バイプロット上で構造状態間の差異を生み出している相互作用を明らかにすることができる. 例えば, 図 3(a)において相互作用 $k = 5, 6, 7, 8$ を強め, 相互作用 $k = 1, 2, 3, 4$ を弱めると状態 C の比率が増加し, 状態 A の比率が減少することが分かる. このことから相互作用 $k = 5, 6, 7, 8$ は状態 C を状態 A に対して相対的に安定化(あるいは状態 A を状態 C に対して相対的に不安定化)している相互作用であると推測できる. なお, 上記のようにバイプロット上で左側の相互作用を強めると左側の構造比率が増加するというよう

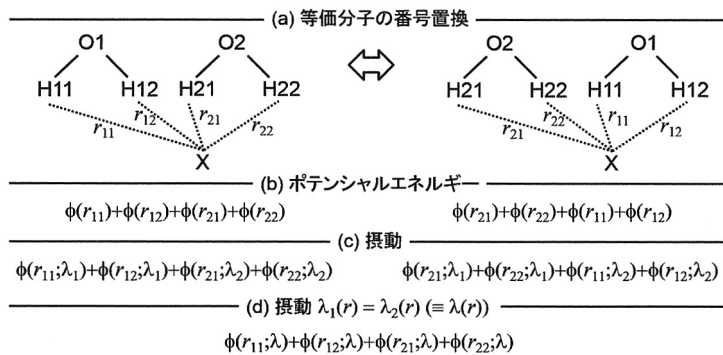


図 4. 等価分子の番号置換に対するポテンシャルエネルギーの不変性と摂動による影響.

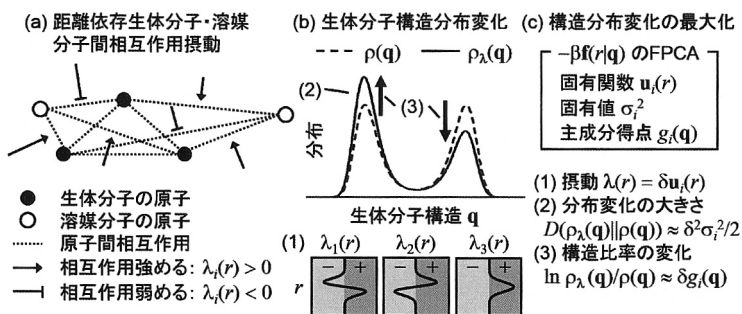


図 5. 距離に依存した生体分子・溶媒分子間相互作用の摂動解析 (DIPA).

な解釈をするためには $\beta V(q)$ ではなく $-\beta V(q)$ を用いて PCA を行うことが重要である.

2.4 距離に依存した生体分子・溶媒分子間相互作用の摂動解析 (DIPA)

次に、水やイオンなどの溶媒分子が生体分子に及ぼす影響を明らかにするために、生体分子・溶媒分子間相互作用への摂動解析を考える。まず、2.2 節で導入した相互作用の強弱を変化させる摂動を適用した場合について考える。溶媒分子の場合には分子内の置換対称性だけでなく、溶媒分子の番号入れ替え (図 4(a)) に対するポテンシャルエネルギーの置換対称性が存在する (図 4(b))。このため、摂動後の置換対称性を保つためには置換対称性のある相互作用を全ての溶媒分子についてまとめて一つのポテンシャルエネルギー項として扱う必要があるが、このポテンシャルエネルギー項に対して摂動を加えると、生体分子近くにある溶媒分子も遠くにある溶媒分子も同じように相互作用の強さが変化する。このように、溶媒分子の距離によらず一様に相互作用の強さを変化させる摂動解析 (intermolecular perturbation analysis; IPA; Koyama et al., 2011) では主成分得点による構造状態の分離が不十分であり、溶媒の生体分子に対する距離依存的な寄与を考慮する必要があることが示唆された。

溶媒分子の距離依存的な寄与を考えるためには、距離に応じて相互作用の強弱を変えることができる摂動解析を行えばよい (図 5(a), (b))。原子間距離 r での相互作用のポテンシャルエネルギーを $\phi(r)$ 、距離に依存した摂動パラメーターを $\lambda(r)$ とすると原子間に働く力が $(1 + \lambda(r))$ 倍となる

ような摂動を考えたい. このような摂動に対応するポテンシャルエネルギーは $F(r) \equiv -d\phi(r)/dr$ とすると

$$(2.9) \quad \phi(r; \lambda) \equiv \begin{cases} \int_r^{r_c} \lambda(r') F(r') dr', & r \leq r_c \\ 0, & r > r_c \end{cases}$$

により実現できる. ここで, カットオフ距離 r_c を導入した. r_c の値を大きくしたときに摂動解析の結果が収束するかどうかを確認し, 収束していればカットオフが無い ($r_c = \infty$) 場合の結果を表していると考えられる. また, 摂動パラメーターが距離に依存しない場合と異なり, 摂動項が $\lambda(r)\phi(r)$ とはならないことに注意が必要である. 距離に依存した摂動の場合にも個別に摂動を加えると摂動後の置換対称性が失われるため(図 4(c)), 置換対称性のある相互作用については同一の摂動パラメーターを用いて摂動を加える必要がある(図 4(d)).

以上の結果に基づき, 置換対称性を持つ相互作用のグループ k に対して, 置換対称性を保つ距離依存的な摂動は, 共通の相互作用のポテンシャルエネルギー関数を $\phi_k(r)$, 共通の摂動パラメーターを $\lambda_k(r)$ とすると

$$(2.10) \quad \Delta V_k(\mathbf{q}, \mathbf{q}') \equiv \sum_{i \in I_k} \sum_{j \in J_k} \phi_k(r_{ij}; \lambda_k)$$

で与えられる. I_k は相互作用 k に属する生体分子の等価原子の集合, J_k は相互作用 k に属する溶媒の等価原子の集合, r_{ij} は原子 i と原子 j の間の距離を表す. 次に, 摂動による生体分子構造の分布変化(2.3)にある条件付期待値 $\langle \Delta V_k | \mathbf{q} \rangle$ の評価について考える. 生体分子の構造 \mathbf{q} が固定されているとき, 生体分子の原子 i から距離 r にある J_k に含まれる溶媒原子の平均数密度を $n_{iJ_k}(r|\mathbf{q})$ とすると式(2.10)の条件付期待値は

$$(2.11) \quad \langle \Delta V_k | \mathbf{q} \rangle = \sum_{i \in I_k} \int_0^{r_c} \phi(r; \lambda_k) n_{iJ_k}(r|\mathbf{q}) dr$$

と書ける. ここで, 生体分子の構造 \mathbf{q} が固定されているときの生体分子の原子 i から距離 r 以内にある平均溶媒原子数 $N_{iJ_k}(r|\mathbf{q}) \equiv \int_0^r n_{iJ_k}(r'|\mathbf{q}) dr'$ および

$$(2.12) \quad f_k(r|\mathbf{q}) \equiv F_k(r) \sum_{i \in I_k} N_{iJ_k}(r|\mathbf{q})$$

を導入し, 式(2.11)の部分積分を行うと

$$(2.13) \quad \langle \Delta V_k | \mathbf{q} \rangle = \int_0^{r_c} \lambda_k(r) f_k(r|\mathbf{q}) dr$$

と書けることが分かる.

以上から, 生体分子・溶媒分子間相互作用に対する距離に依存した摂動 $\Delta V(\mathbf{q}, \mathbf{q}') = \sum_{k=1}^M \Delta V_k(\mathbf{q}, \mathbf{q}')$ による構造分布変化の大きさ(2.3)は, $\mathbf{f}(r|\mathbf{q}) = (f_1(r|\mathbf{q}), \dots, f_M(r|\mathbf{q}))^T$ および $\boldsymbol{\lambda}(r) = (\lambda_1(r), \dots, \lambda_M(r))^T$ を用いて

$$(2.14) \quad D(\rho\boldsymbol{\lambda}(\mathbf{q}) || \rho(\mathbf{q})) = \frac{1}{2} \text{var} \left(\beta \sum_{k=1}^M \int_0^{r_c} \lambda_k(r) f_k(r|\mathbf{q}) dr \right) + \dots$$

$$(2.15) \quad = \frac{1}{2} \langle \boldsymbol{\lambda}, \hat{C} \boldsymbol{\lambda} \rangle + \dots$$

と書ける. ここで, 内積 $\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}' \rangle \equiv \int_0^{r_c} \boldsymbol{\lambda}(r)^T \boldsymbol{\lambda}'(r) dr$ および

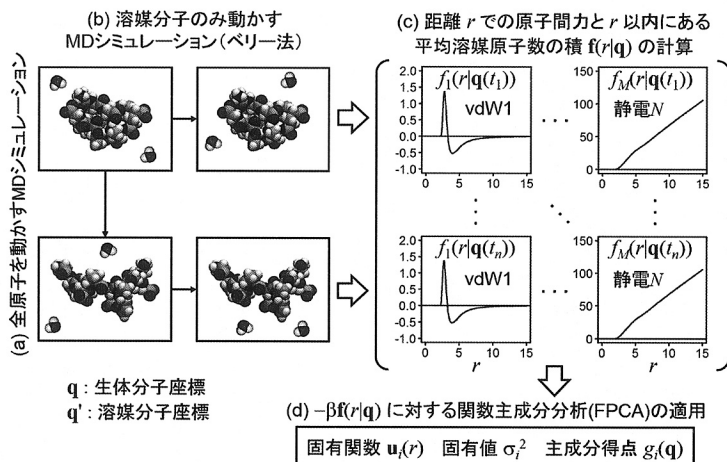


図 6. DIPA 実行手順.

$$(2.16) \quad \hat{C}\lambda(r) \equiv \int_0^{r_c} \text{cov}(\beta f(r|q), \beta f(r'|q)) \lambda(r') dr'$$

のように作用する共分散作用素 (covariance operator) \hat{C} を導入した. $\text{cov}(\beta f(r|q), \beta f(r'|q))$ はその i, j 要素が $\text{cov}(\beta f_i(r|q), \beta f_j(r'|q))$ となるような共分散行列である. これにより生体分子構造の分布を最も大きく変化させるような摂動 $\lambda(r)$ を見つけるためには式 (2.14) の分散を最大化すれば良い. $|\lambda| \equiv \sqrt{\langle \lambda, \lambda \rangle}$ と定義すると, $|\lambda| = \delta$ という条件での式 (2.14) の分散の最大化は関数データ $\beta f(r|q)$ または $-\beta f(r|q)$ を用いた関数主成分分析 (functional PCA; FPCA; Jolliffe, 2002; Ramsay and Silverman, 2006) により実行できる. 従って, 共分散作用素 \hat{C} の降順の固有値を σ_i^2 , 対応する固有関数を $u_i(r)$ とすると $\lambda(r) = \delta u_1(r)$ のとき $|\lambda| = \delta$ という条件で分散が最大になる. 同様に, $\lambda(r) = \delta u_i(r)$ は $|\lambda| = \delta$ および $\langle u_1, \lambda \rangle = 0, \dots, \langle u_{i-1}, \lambda \rangle = 0$ という条件で分散を最大化することが分かる. 摂動 $\lambda(r) = \delta u_i(r)$ による構造分布変化の大きさ (2.15) は式 (2.7) となり, 構造比率の変化 (2.2) は $g_i(q) \equiv -\beta \langle u_i(r), f(r|q) - \langle f(r|q) \rangle \rangle$ と定義すると式 (2.8) となる. $g_i(q)$ は定義から $-\beta f(r|q)$ を用いた FPCA の第 i 主成分得点である. 以上から, $-\beta f(r|q)$ の FPCA を行うことで距離に依存した生体分子・溶媒分子間相互作用の摂動解析 (distance-dependent intermolecular perturbation analysis; DIPA; Koyama et al., 2011) が実行できることが分かる (図 5(c)).

DIPA を実行するには平均溶媒原子数 $N_{iJ_k}(r|q)$ を推定する必要がある. このためには, 生体分子を固定して, 溶媒分子のみを動かす MD シミュレーションを実行する必要があるが, このような特定の自由度だけを動かす MD シミュレーションの方法はベリー (belly) 法として知られている (古明地 他, 2000; Case et al., 2008). また, 生体分子構造 q は平衡状態からサンプリングする必要がある. 以上から, DIPA の実行手順は次のようになる. まず, 生体分子と溶媒分子が両方とも動く通常の全原子 MD シミュレーションを実行する (図 6(a)). この結果得られた生体分子構造を適当な間隔で初期構造として, ベリー法による溶媒分子のみを動かす複数の MD シミュレーションを実行する (図 6(b)). 各シミュレーションで平均溶媒原子数 $N_{iJ_k}(r|q)$ を推定後, 式 (2.12) を用いて $f_k(r|q)$ を計算し (図 6(c)), 関数データ行列 $-\beta f(r|q)$ に対する FPCA を実行すればよい (図 6(d)).

3. シニョリンフォールディングシミュレーションの相互作用摂動解析

10個のアミノ酸(Gly1-Tyr2-Asp3-Pro4-Glu5-Thr6-Gly7-Thr8-Trp9-Gly10)からなり、 β ヘアピン構造に折り畳まれるシニョリン(chignolin; Honda et al., 2004)のフォールディングシミュレーションを用いて、2.2節で導入した生体分子内相互作用の摂動解析であるPEPCAと2.4節で導入した距離に依存した生体分子・溶媒分子間相互作用の摂動解析であるDIPAを適用し、バイプロットを用いて結果の解釈を行う。

3.1 生体分子内相互作用の摂動解析(PEPCA)の適用

シニョリンのNMR構造(PDB ID: 1UAO)の1番目の構造に対して、シニョリンから20Å以内が水に満たされるように、水分子6437個とイオン(12Na⁺と10Cl⁻)を追加した。ポテンシャルエネルギーの極小化を行った後、100ピコ秒の温度・圧力(1気圧)一定(NpT)シミュレーションを行い、1気圧の密度に体積を緩和させた結果、一辺が58.4Åの立方体となった。この後、1マイクロ秒の温度・体積一定(NVT)シミュレーションを行った。いずれのシミュレーションもシニョリンの融解温度である39°C(Honda et al., 2004)で行った。MDシミュレーションはAMBER10(Case et al., 2008)によりLangevin方程式と周期境界条件を用いて実行した。シニョリンは138原子からなり、分子内相互作用数は18,919であるが、置換対称性を保つようにグループ化すると $M = 12,927$ となった。1マイクロ秒の構造データから1ナノ秒ごとに置換対称性のあるポテンシャルエネルギーを計算し($n = 1000$ データ)、 $M \times n$ 中心化データ行列の特異値分解(SVD)を行うことでPCAを実行した(付録A)。

PEPCAの固有値(図7)は第1, 2固有値が不連続的に大きく、第3固有値以降は連続的になっており、第1, 2固有ベクトルでシニョリンの構造状態の安定性を決定している相互作用が特徴付けられることが期待される。第1, 2主成分得点の散布図(図8:点)を見ると、大きく分けて4つの状態があることが分かる。このうち、左側の状態に対応する構造は β ヘアピン構造であることから「正しく折り畳まれた状態」(Native; N)、右上の状態に対応する構造は「ほどけた状態」(Unfold; U)と呼ぶことにする。一方、シニョリンのマルチカノニカルMDシミュレーションの結果によりNative状態である β ヘアピン構造ではAsp3H(本稿ではバイプロットでの表記上、3Hなどと省略する)がThr8Oと水素結合しているが、Asp3HがGly7Oと水素結合したMisfold状態が存在することが見出された(Satoh et al., 2006)。右下の状態に対応する2つの構造はこのMisfold状態に相当するため「誤って折り畳まれた状態1, 2」(Misfold 1, 2; M1, M2)と呼ぶことにする。

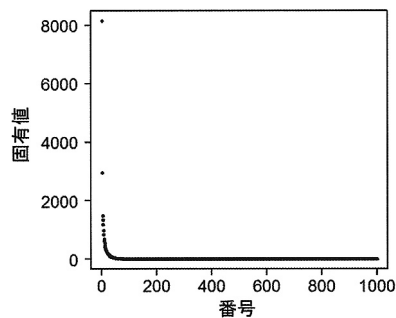


図7. PEPCA固有値(Koyama et al., 2011).

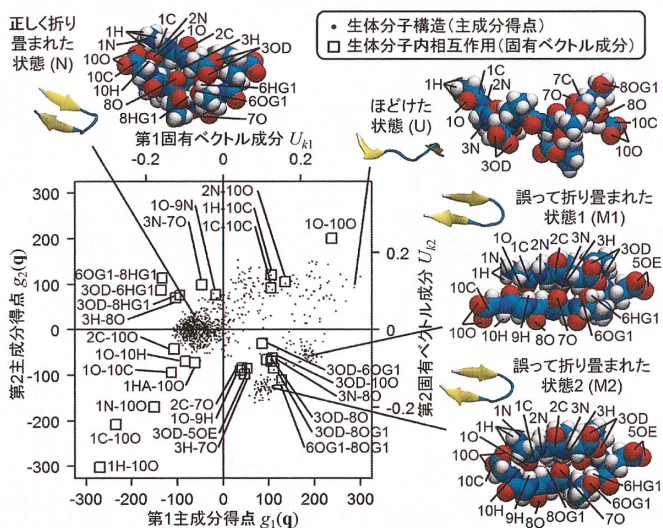


図 8. PEPCA バイプロット (Koyama et al., 2011). 6OG1-8HG1 などは 6 番目のアミノ酸 (トレオニン) のヒドロキシル基の酸素原子 (O) と 8 番目のアミノ酸 (トレオニン) のヒドロキシル基の水素原子 (H) との静電相互作用を表す。小さい点は第 1, 2 主成分得点, 小さい四角は第 1, 2 固有ベクトル成分の散布図を表す。第 1 または第 2 固有ベクトル成分の絶対値が上位 20 位までに入る相互作用のみを示す。PEPCA の解析には二面角やファンデルワールス相互作用のポテンシャルエネルギーも含むが, 図の相互作用は全て静電相互作用である。また, 立体構造において見やすくするために 9 番目のアミノ酸 (トリプトファン) の側鎖は表示していない。

2.3 節の議論に基づき, この 4 状態の差異を生み出している相互作用は主成分得点の散布図 (図 8: 点) と固有ベクトル成分の散布図 (図 8: 四角) を同時に表示したバイプロットを用いて理解できる。N 状態と M1+M2 状態は左上から右下方向の軸で最もよく分離できる。このため, これらの状態の差異を生み出している相互作用は左上と右下の相互作用を見ればよい。N 状態を安定化あるいは M1+M2 状態を不安定化している相互作用は左上にある相互作用を見ればよく, 左上にある 6OG1-8HG1 や 3OD-6HG1 という相互作用は酸素原子 O が負の部分電荷, 水素原子 H が正の部分電荷を持つため, 引力的な静電相互作用 (静電引力) である。従って, これらの相互作用は N 状態を安定化していることが推測できる。実際に, N 状態の構造 (図 8 の vdW 球表示。以下では N 構造などとする。) で 6OG1 と 8HG1 や 3OD と 6HG1 の間の相互作用が確認できるが, M1 や M2 の構造では見られないことが分かる。また, M1+M2 状態を安定化あるいは N 状態を不安定化している相互作用は右下の相互作用を見ればよい。右下の 3OD-6OG1 などが集まっている相互作用は全て静電斥力であり, N 状態を不安定化していることが示唆される。右下にある 3H-7O は静電引力であり, M1+M2 状態を安定化していることが示唆される。実際に, M1 と M2 の構造上で 3H と 7O の相互作用が確認できるが, N 構造では 3H は 7O ではなく 8O と相互作用していることが分かる。

N+M2 状態と U+M1 状態は左下から右上方向の軸で分離でき, その差異は左下と右上の相互作用を見ればよい。左下の静電引力 1H-10O は N+M2 状態を安定化していることが示唆されるが, 実際に N 構造や M2 構造では 1H と 10O が相互作用しているのに対して, U 構造や M1

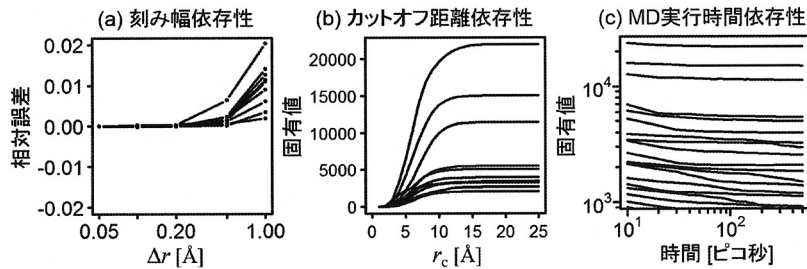


図9. DIPA 固有値の計算条件依存性. (a)第1-10固有値の刻み幅 Δr 依存性. $\Delta r = 0.05 \text{ \AA}$ での固有値に対する相対誤差. $r_c = 25 \text{ \AA}$ を用いた. (b)第1-10固有値のカットオフ距離 r_c 依存性. $\Delta r = 0.05 \text{ \AA}$ を用いた. (a)と(b)では100ピコ秒のベリー法を用いたMDシミュレーションの結果を用いた. (c)固有値のベリー法を用いたMDシミュレーションの実行時間依存性(Koyama et al., 2011). $\Delta r = 0.05 \text{ \AA}$ および $r_c = 25 \text{ \AA}$ を用いた.

構造では離れていることが確認できる. 以上のようにして, 生体分子内相互作用の摂動解析であるPEPCAの結果をパイプロットを用いて表示することで, 主成分得点により生体分子の構造状態の同定と, 固有ベクトル成分により状態間の差異を生み出している生体分子内相互作用を詳細に明らかにすることができる. また, 各状態間に対する安定化あるいは不安定化の寄与は, 静電相互作用が引力的であるかまたは斥力的であるかにより判別できる.

3.2 距離に依存した生体分子・溶媒分子間相互作用の摂動解析(DIPA)の適用

次に, 距離に依存した生体分子・溶媒分子間相互作用の摂動解析であるDIPAを用いた解析を行う. 前節で用いた1マイクロ秒のデータから1ナノ秒ごとの構造を初期構造として(図6(a))ベリー法を用いた溶媒分子(水, Na^+ , Cl^-)のみを動かす500ピコ秒のMDシミュレーションを合計 $n = 1000$ 本行った(図6(b)). 平均溶媒原子数は周期境界の隣接する領域も考慮に入れて推定し, $f_k(r|q)$ を計算した(図6(c)). シニョリンと溶媒の置換対称性を考慮した相互作用の数は $M = 798$ であった. FPCA(図6(d))は離散化した関数データに対するPCAにより行った(付録B). カットオフ距離 $r_c = 25 \text{ \AA}$ と距離 r の離散化の刻み幅 $\Delta r = 0.05 \text{ \AA}$ を用いた場合には, 離散化したデータの変数の数は $p = Mr_c/\Delta r = 399,000$ と大きく, SVDを用いたPCAが実行できなかった. 一方で $n = 1000$ と小さいため, $n \times n$ 中心化グラム行列の対角化が可能であり, これを用いてPCAを実行した(付録A).

まず, DIPAを実装するときに必要なパラメーターの値を変えたときに固有値がどのように変化するかを調べた. 図9(a)はFPCAを離散化したPCAで推定するときに導入した刻み幅 Δr 依存性である. $\Delta r \leq 0.2 \text{ \AA}$ であれば固有値が収束しており, 離散化の影響が無くなることが分かる. 図9(b)は式(2.9)で導入したカットオフ距離 r_c 依存性である. $r_c \geq 20 \text{ \AA}$ で固有値が収束しており, カットオフによる影響が無くなることが分かる. 図9(c)は平均溶媒原子数 $N_{i,j,k}(r|q)$ を推定するために必要な溶媒分子のみ動かすMDシミュレーションの実行時間依存性である. 100ピコ秒ほど実行すれば第1, 2固有値は収束しており, 第3固有値もほぼ収束していることが分かる. 以上のように, 十分な精度のパラメーターを用いればその値に依存しない, DIPA固有の結果が得られることが分かる. 以下の解析では, $\Delta r = 0.05 \text{ \AA}$, $r_c = 25 \text{ \AA}$, 100ピコ秒のMDシミュレーションでの結果を用いた.

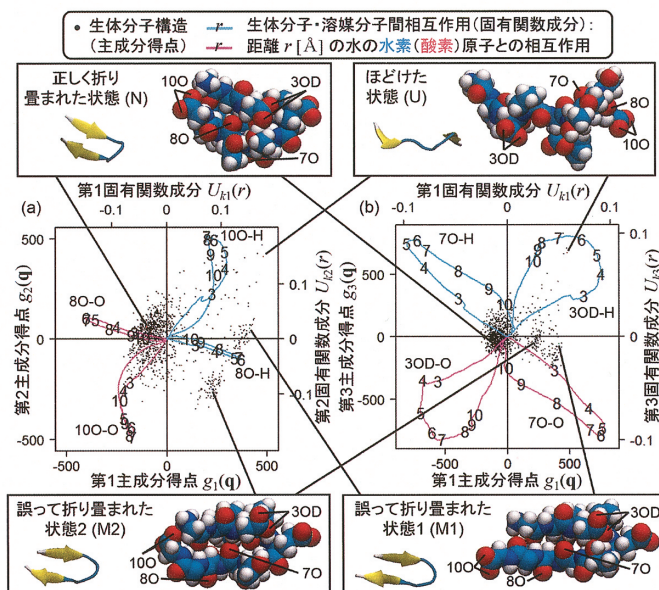


図 10. DIPA バイプロット (Koyama et al., 2011). 小さい点は主成分得点, 曲線は固有関数成分であり, 曲線上の数字はシニョリン原子からの距離 [Å] を表す. 100-O などとはシニョリンの 10 番目のアミノ酸の酸素原子 (O) と水の酸素原子 (O) との静電相互作用を表す. 第 1, 2, 3 固有関数成分の寄与が上位 5 位までに入る相互作用の一部を示す. ここに示していない相互作用のバイプロットや固有関数成分の寄与を評価する計算法については Koyama et al. (2011) にある.

図 9(b), (c)を見ると, 第 1-3 固有値が大きいことから第 1-3 固有関数成分により, シニョリンの構造状態に影響を及ぼす溶媒との相互作用が得られることが期待される. PEPCA の結果と同様に, 第 1, 2 主成分得点の散布図 (図 10(a) : 点) は 4 つの状態を分離できていることが分かる. また, 第 3 主成分得点は N+M1+M2 状態と U 状態を分離していることが分かる (図 10(b) : 点).

次に, 各状態の差異を生み出している溶媒との相互作用を明らかにするためにバイプロットを用いて解析を行う. 2.3 節でのバイプロットについての議論は固有ベクトル成分を固有関数成分に置き換えることで DIPA でもそのまま成り立つが, 固有関数成分は距離 r をパラメータとして連続的に変化するため, 曲線で表される. 第 1, 2 主成分のバイプロット (図 10(a)) から N 状態と M1+M2 状態を識別している相互作用は 80-O または 80-H であり, 静電引力 80-H が M1+M2 状態を安定化していることが示唆される. M1, M2 構造を見ると, 80 原子は溶媒に露出しているのに対して, N 構造ではシニョリン内部に埋もれており, 水による 80 原子の安定化は M1+M2 状態の方が N 状態より大きいことが確認できる. また, 固有関数成分の値は 6Å あたりで最大となっていることが分かる. 距離が離れていくと個々の水との相互作用が弱まっていく効果と, 水の個数が増えていく効果があるが, この両者を合わせた効果の構造変化による差が, 6Å 付近で最大となるためと考えられる. 図 10(b) から静電引力 70-H が M1+M2 状態に対して N 状態を安定化していることが示唆されるが, N 構造では 70 原子が溶媒に露出しているのに対して, M1, M2 構造では 70 原子はシニョリン内部に埋もれていることが確認できる.

同様に、M1とM2の相違を生み出している相互作用は10O-Hと10O-Oであり、静電引力10O-HによりM1が安定されることが示唆される(図10(a))。実際にM1構造では10Oは外に露出しているのに対して、M2構造では内部で相互作用することで、露出が少なくなっていることが分かる。第1,3主成分のパイプロット(図10(b))を見ると、U状態とN+M1+M2状態の差異を生み出している分子間相互作用は3OD-Hと3OD-Oであり、静電引力3OD-HによりU状態が安定化されていることが分かる。実際にU状態の構造を見ると3ODは外に露出しているのに対して、N, M1, M2の構造では3ODは露出が少なくなっていることが確認できる。以上のように、DIPAでは生体分子の構造変化による特定の原子の周りの溶媒和状態の変化を識別することで構造状態を分離し、その差異を生み出している距離依存的な溶媒との相互作用を明らかにすることが可能である。

4. おわりに

本稿では生体分子機械を理解するために相互作用の強弱を変える摂動を導入し、その応答性を生体分子構造の分布変化で定量化する摂動解析について紹介を行った。この結果として、生体分子内相互作用に対する摂動解析としてPEPCAを導入し、距離に依存した生体分子・溶媒分子間の相互作用に対する摂動解析としてDIPAを導入した。このとき、生体分子構造の分布を大きく変化させるような相互作用の摂動は固有ベクトル/関数、摂動による構造分布変化の大きさは固有値、摂動による構造の比率変化は主成分得点、で得られた。また、これらの結果をパイプロットで表示することにより、構造状態とその差異を生み出している相互作用を明らかにできることを示した。

付 録

A. PCAの異なる実装法

MDシミュレーションで一般的に用いられている原子位置のPCA(Kitao et al., 1991; García, 1992; Amadei et al., 1993)や二面角を用いたPCA(dihedral angle PCA; dPCA; Mu et al., 2005; Altis et al., 2007)では、通常は変数の次元 p が数千以下となり、共分散行列の対角化を行うことが可能であることが多い。これに対して本稿で述べたポテンシャルエネルギー項を用いたPCAやFPCAを推定するためのPCA(付録B)では数万以上の変数となり共分散行列の対角化を行うことが困難であることが多い。ここでは共分散行列の対角化を用いないPCAの実装法について簡単に述べる。

構造データとして n 個のデータ $\mathbf{q}_1, \dots, \mathbf{q}_n$ が得られたとする。それぞれのデータについてポテンシャルエネルギーや離散化した関数データなどの統計量 $f_i(\mathbf{q})$ を計算した後、平均を差し引いた(中心化した)統計量

$$(A.1) \quad \bar{f}_i(\mathbf{q}) \equiv f_i(\mathbf{q}) - \frac{1}{n} \sum_{j=1}^n f_i(\mathbf{q}_j), i = 1, \dots, p$$

からなるデータ行列 $\bar{\mathbf{X}}_{ij} \equiv \bar{f}_i(\mathbf{q}_j)$ を考える。共分散行列 $\mathbf{C} = \bar{\mathbf{X}} \bar{\mathbf{X}}^T / n$ の対角化 $\mathbf{C} = \sum_{i=1}^r \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^T$ が実行できる場合にはその固有値と固有ベクトルを計算した後、主成分得点は $\bar{\mathbf{X}}^T \mathbf{u}_i$ と計算できる。

$\bar{\mathbf{X}}$ は特異値分解(singular value decomposition; SVD)により特異値 $\lambda_i > 0$ 、左特異ベクトル \mathbf{u}_i 、右特異ベクトル \mathbf{v}_i を用いて

表 1. PCA の固有値, 固有ベクトル, 主成分得点の計算方法.

| | | | |
|--------------|---|---|--|
| 分解する行列 | 共分散行列 $\mathbf{C} = \frac{1}{n} \bar{\mathbf{X}} \bar{\mathbf{X}}^T$ | 中心化データ行列 $\bar{\mathbf{X}}$ | 中心化グラム行列 $\bar{\mathbf{G}} = \bar{\mathbf{X}}^T \bar{\mathbf{X}}$ |
| 行列サイズ | $p \times p$ | $p \times n$ | $n \times n$ |
| 行列成分 | $C_{ij} = \text{cov}(f_i, f_j)$ | $\bar{X}_{ij} = \bar{f}_i(\mathbf{q}_j)$ | $\bar{G}_{ij} = \bar{\mathbf{f}}(\mathbf{q}_i) \cdot \bar{\mathbf{f}}(\mathbf{q}_j)$ |
| 固有値分解, 特異値分解 | $\mathbf{C} = \sum_{i=1}^r \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^T$ | $\bar{\mathbf{X}} = \sum_{i=1}^r \lambda_i \mathbf{u}_i \mathbf{v}_i^T$ | $\bar{\mathbf{G}} = \sum_{i=1}^r \lambda_i^2 \mathbf{v}_i \mathbf{v}_i^T$ |
| 共分散行列固有値 | σ_i^2 | λ_i^2/n | λ_i^2/n |
| 共分散行列固有ベクトル | \mathbf{u}_i | \mathbf{u}_i | $\bar{\mathbf{X}} \mathbf{v}_i / \lambda_i$ |
| 主成分得点 | $\bar{\mathbf{X}}^T \mathbf{u}_i$ | $\lambda_i \mathbf{v}_i$ | $\lambda_i \mathbf{v}_i$ |

$$(A.2) \quad \bar{\mathbf{X}} = \sum_{i=1}^r \lambda_i \mathbf{u}_i \mathbf{v}_i^T$$

と表すことができる. このとき共分散行列は

$$(A.3) \quad \mathbf{C} = \sum_{i=1}^r \frac{\lambda_i^2}{n} \mathbf{u}_i \mathbf{u}_i^T$$

となり, 共分散行列の固有値と固有ベクトルがそれぞれ中心化データ行列の特異値と左特異ベクトルで表されることが分かる. また, 主成分得点は

$$(A.4) \quad \bar{\mathbf{X}}^T \mathbf{u}_i = \lambda_i \mathbf{v}_i$$

となり, 特異値と右特異ベクトルで表されることが分かる. 以上から中心化データ行列の SVD により PCA の固有値, 固有ベクトル, 主成分得点が計算でき, バイプロットに必要な全ての情報が得られることが分かる (Jolliffe, 2002). 統計解析ソフトの R では prcomp 関数は中心化データ行列の SVD, princomp 関数は共分散行列の対角化により PCA を実行している (R Core Team, 2013).

また, p が大きく, SVD や共分散行列の対角化ができない場合でも n が小さく中心化グラム行列 $\bar{G}_{ij} = \bar{\mathbf{f}}(\mathbf{q}_i) \cdot \bar{\mathbf{f}}(\mathbf{q}_j)$ の対角化

$$(A.5) \quad \bar{\mathbf{G}} = \bar{\mathbf{X}}^T \bar{\mathbf{X}} = \sum_{i=1}^r \lambda_i^2 \mathbf{v}_i \mathbf{v}_i^T$$

が計算できる場合にはその固有値 λ_i^2 と固有ベクトル \mathbf{v}_i により, 共分散行列の固有値 λ_i^2/n と主成分得点 $\lambda_i \mathbf{v}_i$ が計算でき, 共分散行列の固有ベクトルは $\mathbf{u}_i = \bar{\mathbf{X}} \mathbf{v}_i / \lambda_i$ と計算できることが分かる (Bishop, 2006). 以上の結果をまとめたのが表 1 である.

式 (A.1) を用いると中心化グラム行列の要素 \bar{G}_{ij} はグラム行列の要素 $G_{ij} = \mathbf{f}(\mathbf{q}_i) \cdot \mathbf{f}(\mathbf{q}_j)$ を用いて

$$(A.6) \quad \bar{G}_{ij} = G_{ij} - \frac{1}{n} \sum_{l=1}^n G_{lj} - \frac{1}{n} \sum_{m=1}^n G_{im} + \frac{1}{n^2} \sum_{l,m=1}^n G_{lm}$$

と書き下せることが分かる. このため, データの中心化をあらかじめ行わなくても, グラム行列を求めた後, 式 (A.6) により中心化グラム行列を計算し, その対角化を実行すれば共分散行列の固有値と主成分得点が計算できることが分かる. カーネル PCA (Schölkopf et al., 1998) はカーネル $k(\mathbf{q}, \mathbf{q}') = \mathbf{f}(\mathbf{q}) \cdot \mathbf{f}(\mathbf{q}')$ を用いることで $\mathbf{f}(\mathbf{q})$ を経由することなくグラム行列を計算し,

中心化グラム行列の対角化により PCA を実行する. カーネル法分野では $\mathbf{f}(\mathbf{q})$ は特徴空間 (feature space) と呼ばれる. カーネル PCA では $\mathbf{f}(\mathbf{q})$ を直接計算しないため, ガウシアンカーネル $k(\mathbf{q}, \mathbf{q}') = \exp(-|\mathbf{q} - \mathbf{q}'|^2/2\sigma^2)$ のように特徴空間が無限次元の場合も扱えるという利点がある.

B. FPCA の実装

ここでは FPCA の最も簡単な実装である関数データの離散化を用いた方法を説明する (Jolliffe, 2002; Ramsay and Silverman, 2006). FPCA では関数データ $\mathbf{f}(r|\mathbf{q}) \equiv (f_1(r|\mathbf{q}), \dots, f_M(r|\mathbf{q}))^T$ に対して線形結合 $\sum_{k=1}^M \int_0^{r_c} \lambda_k(r) f_k(r|\mathbf{q}) dr$ の分散が最大化になるような係数 $\lambda(r) \equiv (\lambda_1(r), \dots, \lambda_M(r))^T$ を探索する. 積分区間 $[0, r_c]$ を N_B 個の区間に離散化し, $\Delta r \equiv r_c/N_B$ および $r_l \equiv l\Delta r$ とすると

$$(B.1) \quad \sum_{k=1}^M \int_0^{r_c} \lambda_k(r) f_k(r|\mathbf{q}) dr \approx \sum_{k=1}^M \sum_{l=1}^{N_B} \lambda_k(r_l) f_k(r_l|\mathbf{q}) \Delta r$$

のように近似できる. これを MN_B 個の係数 $\lambda_k(r_l)\sqrt{\Delta r}$ と変数 $f_k(r_l|\mathbf{q})\sqrt{\Delta r}$ の線形結合とみなし, 通常の PCA を適用した結果の固有ベクトルを $\mathbf{u}_i(r_l)\sqrt{\Delta r}$, 固有値を σ_i^2 , 主成分得点を $g_i(\mathbf{q})$ とする. このとき, PCA の固有ベクトルの直交条件 $\sum_{l=1}^{N_B} (\mathbf{u}_i(r_l)\sqrt{\Delta r})^T \mathbf{u}_j(r_l)\sqrt{\Delta r} = \delta_{i,j}$ は $\Delta r \rightarrow 0$ とすれば, FPCA の固有関数の直交条件 $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{i,j}$ に収束することが分かる. また, PCA の主成分得点

$$(B.2) \quad g_i(\mathbf{q}) = \sum_{l=1}^{N_B} (\mathbf{u}_i(r_l)\sqrt{\Delta r})^T (\mathbf{f}(r_l|\mathbf{q})\sqrt{\Delta r} - \langle \mathbf{f}(r_l|\mathbf{q})\sqrt{\Delta r} \rangle)$$

も $\Delta r \rightarrow 0$ とすれば, FPCA の主成分得点 $g_i(\mathbf{q}) = \langle \mathbf{u}_i(r), \mathbf{f}(r|\mathbf{q}) - \langle \mathbf{f}(r|\mathbf{q}) \rangle \rangle$ に収束することが分かる. PCA および FPCA の固有値は主成分得点の分散であることから, $\Delta r \rightarrow 0$ により PCA の固有値も FPCA の固有値に収束することが分かる. 以上のように, 関数データを離散化後, $\mathbf{f}(r_l|\mathbf{q})\sqrt{\Delta r}$ に対して PCA を実行すれば, その固有値と主成分得点そのまま FPCA の対応する推定値となる. また, PCA の固有ベクトル $\mathbf{u}_i(r_l)\sqrt{\Delta r}$ を $\sqrt{\Delta r}$ で割ることにより, FPCA の固有関数 $\mathbf{u}_i(r)$ が推定できる.

謝 辞

本研究は小林徹也博士, 上田泰己博士, 泰地真弘人博士, 友田修司博士, 高野光則博士との議論やサポートがなければ遂行することができませんでした. ここに感謝を致します. 本稿は日本学術振興会多国間国際研究協力事業「次世代計算機・次世代アルゴリズムによる巨大生体分子システムの動的モデル化」の助成を受けたものです. また, 数値シミュレーションの実行は, 理化学研究所情報基盤センターの RIKEN Integrated Cluster of Clusters (RICC) システムを用いて行われました.

参 考 文 献

- Altis, A., Nguyen, P. H., Hegger, R. and Stock, G. (2007). Dihedral angle principal component analysis of molecular dynamics simulations, *Journal of Chemical Physics*, **126**, 244111.
- Amadei, A., Linssen, A. B. M. and Berendsen, H. J. C. (1993). Essential dynamics of proteins, *Proteins*, **17**, 412–425.

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer, New York.
- Case, D. A., Darden, T. A., Cheatham, T. E., III, Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Crowley, M., Walker, R. C., Zhang, W., Merz, K. M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossváry, I., Wong, K. F., Paesani, F., Vanicek, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V. and Kollman, P. A. (2008). AMBER 10, University of California, San Francisco.
- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58**, 453–467.
- García, A. E. (1992). Large-amplitude nonlinear motions in proteins, *Physical Review Letters*, **68**, 2696–2699.
- Gower, J. C. and Hand, D. J. (1996). *Biplots*, Chapman & Hall, London.
- Honda, S., Yamasaki, K., Sawada, Y. and Morii, H. (2004). 10 residue folded peptide designed by segment statistics, *Structure (London)*, **12**, 1507–1518.
- Ikeguchi, M., Ueno, J., Sato, M. and Kidera, A. (2005). Protein structural change upon ligand binding: Linear response theory, *Physical Review Letters*, **94**, 078102.
- Jolliffe, I. T. (2002). *Principal Component Analysis*, 2nd ed., Springer, New York.
- 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.
- 古明地勇人, 上林正巳, 長嶋雲兵 (2000). 生体分子の分子動力学シミュレーション(1)方法, *Journal of Chemical Software*, **6**, 1–36.
- Koyama, Y. M., Kobayashi, T. J., Tomoda, S. and Ueda, H. R. (2008). Perturbational formulation of principal component analysis in molecular dynamics simulation, *Physical Review E*, **78**, 046702.
- Koyama, Y. M., Kobayashi, T. J. and Ueda, H. R. (2011). Perturbation analyses of intermolecular interactions, *Physical Review E*, **84**, 026704.
- Kullback, S. (1959). *Information Theory and Statistics*, John Wiley & Sons, New York.
- Malolepsza, E., Strodel, B., Khalili, M., Trygubenko, S., Fejer, S. N. and Wales, D. J. (2010). Symmetrization of the AMBER and CHARMM force fields, *Journal of Computational Chemistry*, **31**, 1402–1409.
- Ming, D. and Wall, M. E. (2005). Quantifying allosteric effects in proteins, *Proteins*, **59**, 697–707.
- Ming, D. and Wall, M. E. (2006). Interactions in native binding sites cause a large change in protein dynamics, *Journal of Molecular Biology*, **358**, 213–223.
- Mu, Y., Nguyen, P. H. and Stock, G. (2005). Energy landscape of a small peptide revealed by dihedral angle principal component analysis, *Proteins*, **58**, 45–52.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, <http://www.r-project.org/>.
- Ramsay, J. O. and Silverman, B. W. (2006). *Functional Data Analysis*, 2nd ed., Springer, New York.
- Satoh, D., Shimizu, K., Nakamura, S. and Terada, T. (2006). Folding free-energy landscape of a 10-residue mini-protein, chignolin, *FEBS Letters*, **580**, 3422–3426.
- Schölkopf, B., Smola, A. and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, **10**, 1299–1319.
- Zwanzig, R. (2001). *Nonequilibrium Statistical Mechanics*, Oxford University Press, New York.

Perturbation Analyses of Biomolecular Simulations

Yohei M. Koyama

Quantitative Biology Center, RIKEN

To understand protein functions in atomic detail, we need to investigate different main-chain/side-chain conformational states, interaction patterns to other molecules (compounds, DNA, and other proteins), and contributions of solvents (water and ions). We review our recent approaches to identify complex conformational states and their interactions by performing perturbation analyses of atomic interactions. First, we introduce PEPCA (potential energy principal component analysis), which is a perturbation analysis of atomic interactions within proteins or complex. PEPCA is implemented by performing the principal component analysis of interaction energies. PEPCA principal component scores identify conformational states that have different interaction patterns within proteins or complex. PEPCA eigenvector components identify interactions that differentiate each state. Next, we introduce DIPA (distance-dependent intermolecular perturbation analysis), which is a perturbation analysis of atomic interactions between proteins/complex and solvents. DIPA is implemented by performing the functional principal component analysis by using products of intermolecular forces and average numbers of solvent atoms. DIPA principal component scores identify conformational states that have different interaction patterns between proteins/complex and solvents. DIPA eigenfunction components identify interactions that differentiate each state. Finally, we apply PEPCA and DIPA to the chignolin folding simulation. Results are visualized by using biplots that are scatter plots of eigenvector (or eigenfunction) components and principal component scores.