様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月３０日現在

機関番号：１５１０１
研究種目：基盤研究（C）
研究期間：２０１０～２０１２
課題番号：２２５００２７５
研究課題名（和文）　分子種間変動を扱う階層モデルとその正則化による蛋白質立体構造の比較解析
研究課題名（英文）　Comparisons of three-dimensional structures of proteins using hierarchical models and regularization for between-protein variations
研究代表者
網崎　孝志（AMISAKI TAKASHI）
鳥取大学・医学部・教授
研究者番号：２０２３１９９６

研究成果の概要（和文）：蛋白質立体構造の比較解析法を開発した。対象とするデータは複数の蛋白質の三次元座標で、蛋白質それぞれに複数のコンホメーションが得られている場合である。そのような蛋白質 - コンホメーションという階層性に着目し、蛋白質別変動と蛋白質内変動を考慮する混合効果モデルを用いた。推定方法は EM アルゴリズムをベースとしているが、それぞれの蛋白質に固有のズレを推定するために事後分布最大化を用い、さらには、ズレの個所を限定し、分子の特徴を抽出するために L1 正則化を組み入れた。この推定法の各反復において、独自の共分散行列を対象とした特異値分解により、分子の重ね合わせのための回転行列を推定した。数値テストの結果、本比較解析法は蛋白質の特徴的な構造を抽出するために効果的であると考えられた。

研究成果の概要（英文）：A method for comparing protein structures was developed. We consider a collection of three-dimensional coordinates of multiple proteins in which each protein has multiple conformations. The method exploits this protein-conformation hierarchy, and is based on the mixed-effects models on between- and within-protein variability. The estimation procedure is that of the EM-algorithm, in which the individual deviations of proteins are estimated and restricted using maximum-a-posteriori estimators and L1-regularization. At each iteration of the algorithm, the superposition of conformations can be made using a usual singular-value-decomposition but for specialized covariance matrices. The results of numerical tests indicate that the method can be effective for characterizing the structures of multiple proteins simultaneously.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	７００，０００	２１０，０００	９１０，０００
２０１１年度	１，３００，０００	３９０，０００	１，６９０，０００
２０１２年度	１，２００，０００	３６０，０００	１，５６０，０００
年度			
年度			
総　計	３，２００，０００	９６０，０００	４，１６０，０００

研究分野：　総合領域
科研費の分科・細目：　情報学・生体生命情報学
キーワード：　統計数学、生体分子、蛋白質

様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

１．研究開始当初の背景

蛋白質の構造と機能の間の関係の解明は、医学・自然科学における重要課題であり、社会からの期待も大きい。立体構造決定の技術も進歩しており、多数の蛋白質の構造が明らかとなってきているが、残念ながら、構造と機能の間の関係についての普遍的な法則の解明にはいたっておらず、構造から機能を読み取ることは、人の目を頼りにするような段階から抜け出ていない。情報学の手法により、それに取り組む分野が、バイオインフォマティクスの一分野でもある構造バイオインフォマティクスといえる。

一般に、バイオインフォマティクスの最も重要なテーマのひとつとなっているのが遺伝子の機能解明である。その基本は塩基配列を比較し、その相同性から機能を推察することである。同様に、蛋白質の構造と機能についても、基本は、二つの蛋白質の構造を比べることである。このような作業は、X 線回折や NMR などにより実験的に得られた構造だけではなく、分子動力学計算などのコンピュータシミュレーションにより発生された構造の解析でも同様である。

蛋白質の立体構造の比較の基本は、二つのうち、一方を回転・平行移動して、両者の原子ができる限り一致するように重ね合わせることである。通常は、root-mean-squared (rms) 偏差が最小になるように、回転・並進のパラメータを求める。これは、原子位置に等分散性を仮定した最尤法といえる。ただし、等分散性のために、二つの分子が全体的に均一に重なってしまい、違いがよくわからないことがある。これは、ペプチド鎖末端など変動の大きい部分を、他と同じようにフィットしようとすることが原因のひとつである。本研究開始の少し前になって、この問題に対し、末端などゆらぎが大きい部分の重みを小さくとるような異分散性の最尤法が、国外の研究者により提案された。

ところが、その中でも比較的洗練された手法であっても、多くの分野で利用されている反復重みづけ最小二乗法、あるいは、一般化最小二乗法のクラスの手法であった。そこで、より一層洗練された手法を開発することができれば、蛋白質の構造を客観的に、正確に、容易に、迅速に比較することが可能となり、ひいては、蛋白質の機能と構造の関係の解明の一助となればと考え、本研究に着手した。

２．研究の目的

蛋白質構造比較法は、いくつかの種類がある。ひとつは、比較対象の原子の対が既知であるかどうかによる分類である。すなわち、二つの分子を重ね合わせるときにマッチさせる点が、あらかじめ決まっているときの方法と、そのマッチさせる点を決めることから始める方法である。後者は、構造アライメントとよばれることが多い。本研究で取り組む手法は前者に分類される。したがって、比較する蛋白質の構造が比較的似通っているような場合を想定している。また、別の分類としては、構造の変化・変動に物理学的なモデルを仮定するか否かによる分類がある。前者は、たとえば、リガンド結合により、構造が鳥の羽ばたきのように運動するようなモデルを仮定することであり、そのような研究は多数行われている。本研究では、特定の運動などのモデルを仮定することなく、普遍的に構造変化を検出できるような方法、すなわち、統計学的な手法を開発することを目指している。

具体的にいうと、本研究の目的は、階層モデルと L1 正則化を使って、多種の蛋白質の比較を、現状と比べて、容易に、客観的に、精密に、迅速に、効果的に行えるような構造比較法を開発することである。階層モデルを利用するのは、それに Bayes 的な手法を組み合わせることにより、蛋白質の単一種ごとの解析→比較という従来法よりも、全体の情報を有効利用し、それが個々の蛋白質についての結果の精度も向上させることができると考えたからである。また、L1 正則化を利用するのは、個々の蛋白質の構造の特徴を的確につかむには、個々の原子のズレをそのまま評価するのではなく、ズレのある部分を限定してやれば、特徴的な部分が浮かび上がってくると考えたからである。以上のように、本研究の手法は、蛋白質種間で構造変動情報を互いに共同利用することにより不足しがちな情報を補完し（階層モデルと Bayes）、なおかつ、特徴的なズレのある部分を客観的な手法で浮かび上がらせる（L1 正則化）ような構造比較法であり、そのような手法の開発を目指した。

３．研究の方法

（１）推定理論の検討

構造の違いを、蛋白質分子別の変動と蛋白

質分子ごとの分子内変動という二つの要因に分割し、それを扱う階層モデルとして混合効果モデルを利用することとした。そのうえで、理論的考察、実験的考察、文献調査により検討し、各種推定方程式などを作成した。なお、実験的考察とは、プログラムを作成し、実験を行い、その結果を理論面での方法選択や改善に利用したことである。

本手法は、構造の違いを分子別変動と分子内変動という二つの変量効果で説明する混合効果モデルである。分子別変動は蛋白質ごとの特徴を表すものであり、それを限定して抽出するために、分散行列の L1 正則化を用いた。原子単位での選択を可能とするため、三次元行列正規分布を仮定した。これにより、原子位置に異方性の分散を仮定する必要が薄れたため、L1 正則化の実装の容易さも考慮して、分子別分散と分子内分散の両方とも対角行列にとることとした。分子内変動については、当初、同様のあるいは Bayes 的変数選択を想定していたが、分子別変動の選択・抽出が最大の目的であり、分子内変動の推定は、それが効果的に行えればよいと判断し、標本分散行列の正則化による方法に切り替えた。現時点では、Ledoit & Wolf の方法を用いているが、他の固有値縮小の方法も有望と思われる。

（２）プログラムの作成

上記（１）の推定理論を 1500 行程度の C 言語プログラムとして実装した。将来的な、ソースコードでのプログラム頒布を想定し、スクラッチからコーディングを行ったが、現時点では、特異値分解について、LAPACK ルーチンを用いている。

（３）性能検証

比較的分解能が高く、球状でコンパクトな蛋白質として RNase T1(PDBID: 1IOV)を選定し、手動で摂動を加えて複数の「分子種」を作成し、それらに正規乱数を加えてコンホメーション（個々の構造、レプリカ、スナップショット）を発生させた。そのデータを対象に解析を行い、本研究の解析法の性能の評価を行った。

その他、後の性能検証でも利用するために、酸化 GTP 分解酵素のプロトネーションが異なる分子種の分子動力学シミュレーションを行った。

４．研究成果

（１）比較解析法の特徴

本手法の特徴のひとつは、複数の蛋白質分子の比較解析を１度の解析で同時に行えることである。これにより、分子動力学計算のトラジェクトリ解析などでは、構造比較作業に必要な労力が大幅に軽減すると思われる。

また、個々の蛋白質のズレの推定には、事後分布最大化による経験 Bayes 推定量を開発した。このことは、すべての蛋白質の情報を、個々の蛋白質のズレの推定に利用することを意味する。したがって、コンホメーションの標本数の小さい蛋白質についても、その情報を補足するような効果が期待できる。

さらに、個々の蛋白質のズレの位置を限局化するような L1 正則化法を開発し、組み入れた。これにより、それぞれの蛋白質に特徴的なズレを、容易に検出することが可能になると思われる。

（２）数値テスト：全般

RNase T1(PDBID:1IOV)の Cα原子を使った。３個の残基 Gly34, Ser35, Asn36 に手動で摂動（-2Å～+2Å）を加え、「５種類」の蛋白質を作成した。それに正規乱数（標準偏差 1Å）を加え、それぞれの蛋白質に８個のコンホメーションを作成した。それらの同時重ね合わせを行った結果を図１に示す。グレーの細線が各コンホメーションを表し、えんじ色の棒は５種の蛋白質の平均構造を表している。摂動を加えた３残基の部分でのみ各蛋白質にズレがみられ、他の部分では平均構造が一致している。このように特徴的な部分（摂動を加えた部分）を抜き出すことに成功している。

図１：

（３）従来法との比較

従来法とは、５種類の蛋白質のそれぞれに

重ね合わせを行い平均構造を推定し、得られた５個の平均構造を重ね合わせるものである。前項と同様のデータを用いた。ただし、ペプチド両末端各８残基については正規乱数の標準偏差を 8Å、その内側各４残基は 2Åとした。また、蛋白質あたりのコンホメーションを半数の４個とした計算も行った。結果は図２と図３のとおりである。左側の青色が本手法により得られた結果である。末端部分の変動が大きいため、それらを蛋白質内変動としてとらえることができず、部分的に蛋白質別変動として検出しているが、それでも、図の分子の上部の摂動を加えた部分をとらえている。それに対し、グレーで示した従来法では、摂動を加えた部分がやや明確でない。とくに、蛋白質あたりの標本数が少ないと（図３）、その傾向が顕著である。その場合でも、本手法では分子別変動の検出性能の劣化は大きくはない。

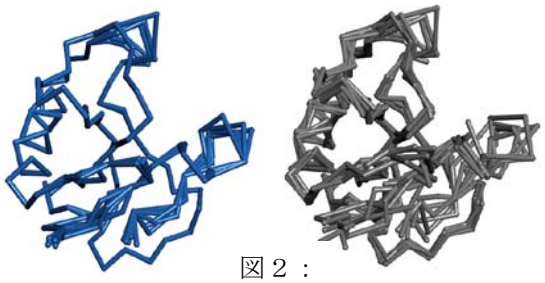

図２：

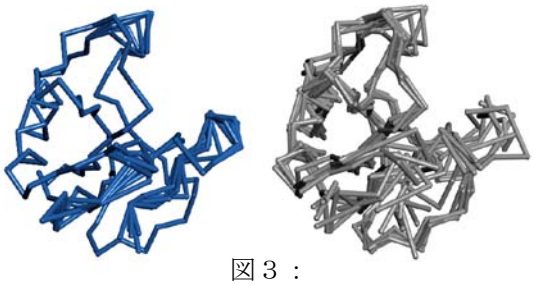

図３：

（４）分散要素の選択

本研究では、L1 正則化の程度と事前分布の寄与度を規定するパラメータλとκを導入した。このうち、λは、それぞれの蛋白質に特有のズレの限局の強さを表している。図２に示したものと同じデータについて、分散要素の個数、すなわち、ズレがあると判定された個数とλの関係を示したものが図４である。赤線はズレありと判定された残基数、緑線は、そのうち、実際に摂動を加えた残基数（正解の個数）である。たとえば、λ=20 では「ズレあり」と判定された残基は４個で、そのうち、実際に摂動を加えた部分は３個である。したがって、蛋白質間でズレのある部分が概ね正確に検出されているといえる。今後、自動的に最適なλの値を求めるような手段の開発が必要と考えている。

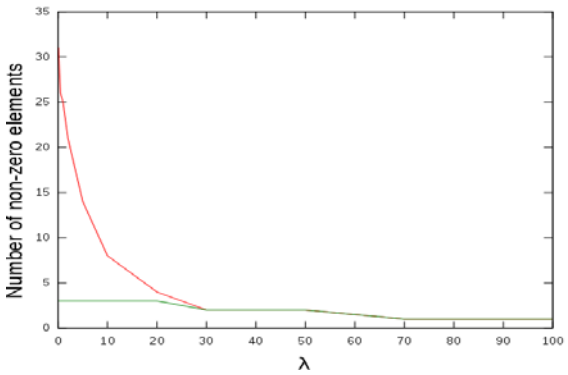

図４：

（５）位置づけ、インパクト、今後の展望

本研究の手法の理論・技術的キーワードは、「階層（混合効果）モデル＋事後分布最大化（Bayes)」と「L1 正則化」である。それぞれは、決して新しいものではないが、この二つを組み合わせて立体構造の比較解析に用いるのは新しいアイデアであり、世界的にもきわめてユニークといえる。本研究では、この二つを組み合わせるための理論を構築し、そして、その二つの効果が数値実験でも確認できた。しかし、本手法の現実的な有効性を明らかにするためには、確率統計学的に十分な検証実験、並びに、現実のデータへの適用による評価が不可欠である。また、λとκの調節方法も検討の必要がある。このように、本手法の普及には、まだ、研究が必要であるが、その原型が開発できたことは、インパクトのある成果だと思われる。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

(1) S. Fujiwara and T. Amisaki, Fatty acid binding to serum albumin: Molecular simulation approaches, Biochim. Biophys. Acta, 査読有, 2013, in press, 10.1016/j.bbagen.2013.03.032

(2) S. Fujiwara and T. Amisaki,
Steric and allosteric effects of fatty acids on the binding of warfarin to human serum albumin revealed by molecular dynamics and free energy calculations,
Chem. Pharm. Bull., 査読有, 59, 2011, 860-867.

〔学会発表〕（計０件）

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
なし

６．研究組織
(1)研究代表者
網崎　孝志(AMISAKI TAKASHI)
鳥取大学・医学部・教授
研究者番号：２０２３１９９６

(2)研究分担者
藤原　伸一（FUJIWARA SHIN-ICHI）
鳥取大学・医学部・講師
研究者番号：００３６２８８０

(3)連携研究者
なし