様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ７ 日現在

機関番号：１５１０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１２２０３

研究課題名（和文）異質性混合効果をベースとした機械学習によるタンパク質ダイナミクス解析法の開発

研究課題名（英文）Methods based on heterogeneous mixed effects model for protein dynamics analysis

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

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：デカルト座標系で表されたタンパク質の立体構造の集まりから、ダイナミクスや構造変動、構造異質性に関する情報を効果的に得るための手法として、立体構造の「集団の集まり」における集団間変動ならびに集団内変動を表す共分散行列を推定するための変量効果モデルに基づく手法を開発した。両変動を分離し、集団間の違いを引き出すことを意図した手法である。また、この「集団の集まり」に対して、異分散性を考慮した重み付き最小二乗法を二段階で適用することにより、同様に、集団内・間共分散行列を推定する手法も開発した。数値実験において、これらの手法で推定した共分散行列から、主要ダイナミクスなど諸量を効果的に推定できた。

研究成果の学術的意義や社会的意義
集団間変動と集団内変動を分離するアイデアは従来よりあったが、デカルト座標系における構造重ね合わせから共分散行列の推定までもを統一した手法はこれまでみられていない。とくに、変量効果モデル法では、他集団の情報をベイズ的に考慮しているので、集団サイズが小さい場合に有効である。このため、データベース上の既知の結晶構造群データからそのタンパク質の構造変動や構造異質性についての知見を得るというような、構造生物学関連での研究で活用されることを期待している。

研究成果の概要（英文）：A method based on a random effects model was developed for estimating covariance matrices that represent the inter- and intra-ensemble variations of protein conformations. The method is applicable under situations such that the dataset is a collection of ensembles each of which is a set of conformations represented in the Cartesian coordinates system. A two-stage method based on the least squares method under heteroscedastic variances was developed as well. Efficiencies of these methods, in particular that of the random effects model at a small ensemble size, were confirmed by numerical tests on machine-synthesized datasets. The estimated principal components for 57 X-ray structures of a kinase were consistent with the functional motifs reported previously. In addition, a combination of the methods with structural clustering was useful for identifying remarkable conformations.

研究分野：構造インフォマティクス

キーワード：変量効果モデル　構造重ね合わせ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

薬物受容体や酵素などタンパク質の機能の詳細の解明は、医学や創薬において重要な課題である。タンパク質の機能は、静的構造だけでなく、薬物受容体の活性型への構造変化のように、ダイナミクスが深く関係している。また、ひとつのタンパク質が、多様な構造をとること自体が機能発現を支えているとも考えられている。このような構造を実験的に明らかにする主要な手法に、X 線結晶構造解析や電子顕微鏡法がある。また、シミュレーションに基づく手法として、分子動力学（MD; molecular dynamics）法は、とくに、ダイナミクスの解析における主要な手法のひとつである。

結晶構造についても、MD 法のトラジェクトリデータを構成する膨大な量の構造についても、そこから構造異質性やダイナミクスを取り出すための解析法が必要である。その場合に、内部座標系による方法とデカルト座標系による方法がある。後者は、結果が直観的に理解しやすいが、構造重ね合わせに起因するアーチファクトという問題点がある。ただ、その程度や性質についての詳細、あるいは、回避法については十分にわかっていない。

研究者は、それまで、混合効果モデルに L1 正則化を導入してタンパク質立体構造の特徴的部分を抽出するような手法を開発していた。

２．研究の目的

本研究では、立体構造の「集団の集まり」という階層性に着目し、主要ダイナミクスの抽出に用いる主成分分析等が対象とする分散共分散行列の推定を、混合効果モデルに基づいて、重ね合わせに起因する問題を解決した上で、デカルト座標上での手法として実現することを目標とした。

３．研究の方法

（１）立体構造の階層モデル

第 i 集団の第 j 構造 Yij について以下の変量効果モデルを仮定する。

$$Yij = M + Zi + Eij,\ j=1,\ldots,mi,\ i=1,\ldots,r \quad (※)$$

r は集団数、mi は第 i 集団のサイズ（構造数）である。M は全集団全構造の母平均構造、Zi と Eij はいずれも確率変数行列で、それぞれ、M からの集団ごとのズレと集団平均 M+Zi からのズレを表す。いずれも、n×3 個のデカルト座標で構成された行列である。本研究では、Zi と Eij はそれぞれ平均 0、共分散行列 V と S の正規分布に従うものとする。それぞれ、集団間変動、集団内変動を表している。実際に観測されているのは Yij に回転 Rij と並進 tij を施した Xij であり、V と S の推定のためには、Rij と tij の推定、すなわち、構造重ね合わせが必要になる。この変量効果モデルに基づき集団間変動を表す共分散行列 V と集団内変動を表す共分散行列 S を推定する手法を整備した。とくに、異方性共分散行列の推定法について考察し、実現した。また、（※）での Zi をパラメータとみなして M+Zi（i=1,...,r）を最小二乗法等により推定し、得られた M+Zi から V を推定することもできる。この場合は（※）は母数効果モデルであり、これに基づく手法も実現した。

（２）生成データを使った数値実験

以上の手法は、正規乱数を用いて発生させた構造に対して適用し、予想された性質についての検証を行った。用いたタンパク質は hMTH1（8-oxo-dGTPase）で、Cα原子 156 個で表された構造を対象とした。生成に使った M、V、S は、各種の条件で実施した MD シミュレーションのトラジェクトリデータから算出したものを用いた。

（３）結晶構造データへの応用

PDB に登録されている分裂促進因子活性化タンパク質キナーゼ（MAPK）p38αの実際の結晶構造での解析を行った。構造は、339 個の Cα原子座標で表現し、それらが揃っていた 59 構造を対象データとした。これに、まず、k=8 の PAM（k-medoids）による構造クラスタリングを行い、集団サイズ 1 のクラスタを除き、集団 6 個、全 57 構造のデータを得た。このデータを用いて、変量効果に基づく手法により共分散行列 V と S を推定した。また、比較目的で、「従来法」により共分散行列を推定した。この行列を D とする。すなわち、全集団の構造をプールして全体を 1 集団とみなし、それに反復重み最小二乗法を適用する手法（PE-IWLS）を用いて推定を行った。

（４）本研究は、タンパク質ダイナミクス解析のための手法の開発が主目的であるが、当初、混合効果モデルに基づく新しい構造クラスタリング法を開発することも目標のひとつとしていた。ところが、混合効果モデルによる同時重ね合わせが予想外に計算時間を要することが判明し、それを繰り返し行うようなクラスタリング法で、実用性に問題があると判断し、開発を見送った。

４．研究成果

（１）二段階法（TS 法）

通常、構造の重ね合わせには最小二乗法（OLS）が用いられる。しかし、タンパク質の構造はコアの部分とそれ以外の末端やループの部分では変動の大きさが一般的に異なる。OLS ではこのことが考慮されておらず、それが、重ね合わせによるアーチファクトの大きな原因のひとつになっていると考えられる。本研究では、そのため、原子ごとに分散が異なることを前提とした異分散性の反復重み最小二乗法（IWLS）の一種を利用した。IWLS（あるいは OLS）により Zi を推定し（母数効果）、r 個の M+Zi (i=1,...,r) に対して再び IWLS（あるいは OLS）を適用して M を推定し、それらを用いて標本共分散行列 V を計算した。共分散行列 S は、第一段階の計算で、個々の集団の標本共分散行列 Si を求め、それに、第二段階で推定した個々の回転行列 Ri を適用し、それらの mi 重み付き平均をもって S の推定値とした。なお、構造重ね合わせの都合で、IWLS の重みにそれぞれの反復での（異方性である）Si の推定値を利用することはできないため、代わりに Si の等方性モデルを作業用分散として用いた。

この TS 法は、集団サイズ mi が大きい場合に有効であることが予想され、実際に、後述の「数値実験」の項で示すように、mi が大きい場合は、とくに、TS-IWLS について、次に述べる REM 法と同程度の精度での推定が可能であった。

（２）変量効果モデルに基づく方法（REM）

M、V、S、およびすべての Yij という全体の情報が与えられれば、それらを使って、Zi の事後分布を知ることができる。これはサイズの小さい集団があったとしても、集団の個数が多ければ、あるいは、他の集団のサイズが大きければ、それらの情報を勘案して、サイズの小さい集団の Zi を効率よく推定できることを意味している。

本研究では、Zi に対するこのベイズ的推定と、変量効果モデル（※）の最尤推定（ML）解に対する EM-アルゴリズムにより、M、V、S、および Rij と tij を推定する手法（REM 法）を実現した。なお、TS-IWLS 法と同様に、REM 法でも異分散性を考慮して Rij を推定する方法とした。また、これも TS-IWLS 法と同様に、構造重ね合わせの都合で、V や S の推定値を作業用分散として利用することはできないため、代わりにそれらに対する等方性モデルを用いた。

タンパク質の立体構造の重ね合わせには、Kabsch の方法に類似の方法が広く使われるが、等方性の場合に限定される。このため、異方性の V や S を最尤法に対する反復法（IWLS や REM の理論的根拠）で推定するのは難しい。単一集団だけを考えるのであれば、TS 法のところで述べたように、作業用分散には等方性のものを用い、重ね合わせが完了したあとで、異方性の、いわゆる標本分散共分散行列を求めることが通常行われる。本研究の REM 法でも、種々検討の結果、同様の考え方で進めた。ただし、作業用分散行列が、V、S ならびに Zi の事後分散の 3 種類存在するために、専用の推定法とした。すなわち、REM 法の EM-アルゴリズムの反復において、重ね合わせのパラメータ（Rij、tij）の値を固定し、分散要素のみの推定を行うものである。種々検討の結果、現時点では、TS 法と条件をそろえるために、母平均構造 M、集団ズレ Zi の推定値も固定し、さらに、反復回数を 1 回だけとする方法を採用した。

（３） 数値実験

集団間と集団内共分散行列（V と S）の対角成分の平方根、すなわち、標本標準偏差の推定値の mean absolute error (MAE) を図 1 に示した。mi や r が増えるほど、MAE は 減少した。いずれの r でも TS-OLS の MAE が他の手法よりも大きかった。mi が小さいところでは、REM 法の MAE が、TS 法のものより小さかった。r が大きくなると、その差が拡大した。

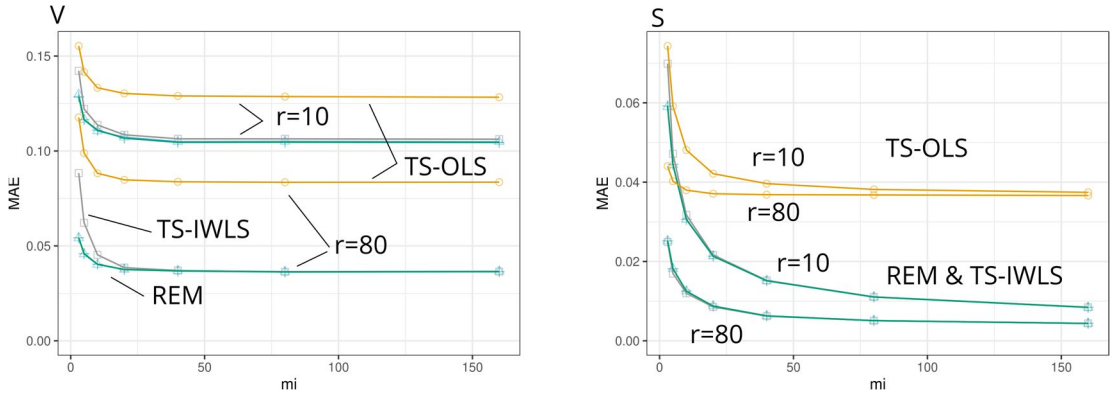

図１：分散の推定誤差

共分散行列の非対角成分については、それを規格化した分相関係数の精度で評価したが、V については r=80、mi=160 とサイズが大きい場合に、TS-IWLS や REM 法で許容できると思われる程度の精度で推定が可能であったが、TS-OLS については、十分な精度での推定はできなかった。S については r や mi が大きい場合に十分な精度で推定できたが、TS-OLS の推定精度は低かった。

固有値と固有ベクトルについては、タンパク質構造のダイナミクスや異質性の解析の場合、上位数個の固有値に対する固有ベクトル(主要ダイナミクスといわれる)が利用されることが多い。そこで、r=10 について上位 9 個、また、r=80 について上位 20 個の固有値の推定精度を方法間で比較した。集団間変動（V）については、分散の場合と異なり、TS-OLS が一様に大きな MAE を与えることはなかった。むしろ、集団サイズが小さい（r=10）場合に、TS-IWLS の推定精度が低かった。集団内変動（S）の固有値については、mi が大きいところで、TS-OLS の MAE が他のものより大きかったが、その程度はわずかであった。概して、REM は、V と S の両方で MAE が他の手法よりも高くなることはなく、また、mi が小さな領域では、他の手法よりも小さい MAE を示し、有効性が確認できた。

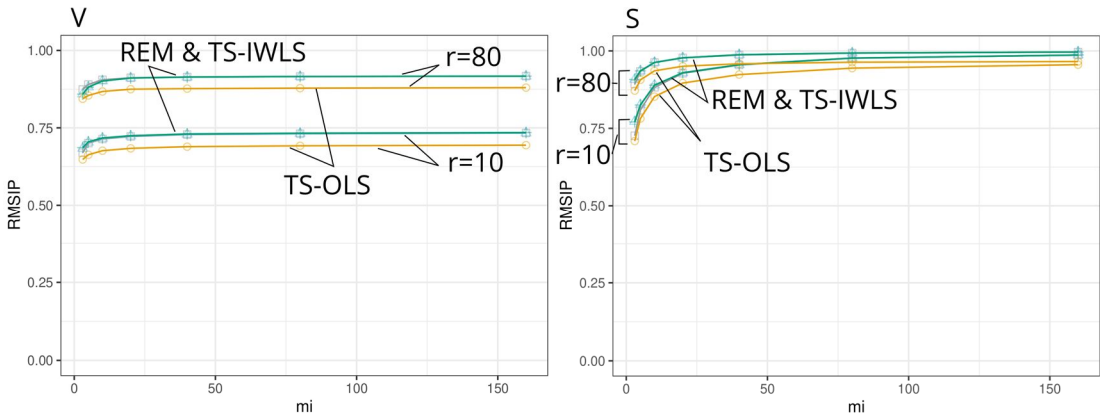

図２：固有ベクトルの推定精度

主要ダイナミクス解析では、特に、上位の少数の固有ベクトルの推定精度が高いことが重要である。図２に r=10 および r=80 について、推定固有ベクトルとその真値との RMSIP を示した。RMSIP は、2 セットの固有ベクトルがそれぞれ張る空間のオーバーラップを表しており、セット間ですべてのベクトル内積の二乗平均の平方根をベクトル数で除したものである（その値が 1 の場合は完全に一致することを意味する）。r=10 の場合は上位 9 個、r=80 については上位 20 個についての値を示した。いずれも高い値を示している。REM と TS-IWLS の精度に大きな違いはなかった。期待された mi が小さいところでの REM の優位性はわずかであった。TS-OLS の値は、全般的にやや低くかった。

（４）p38αの結晶構造への応用

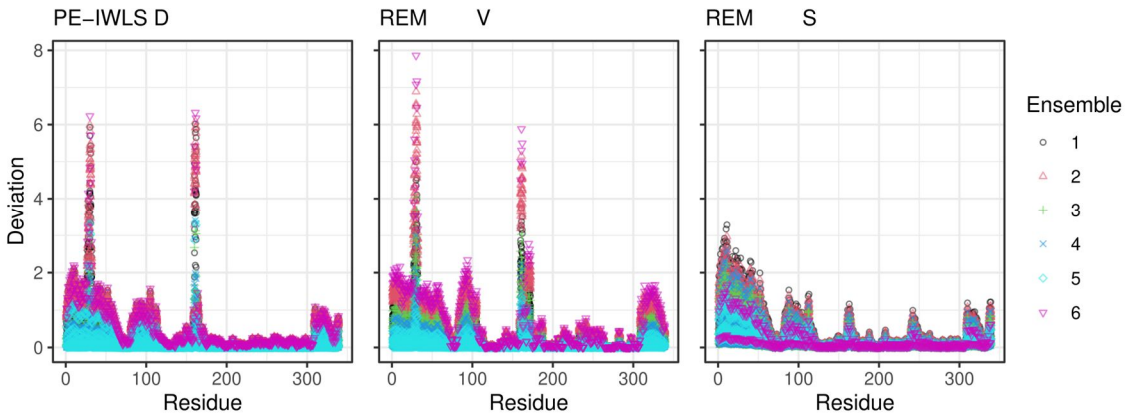

図３：共分散行列の第一主成分（固有ベクトル成分）をデカルト座標に射影し、残基ごとの大きさを示したもの。

図３は、第一主軸について、それぞれの構造の成分をデカルト座標上に射影し、各残基（Cα原子）の大きさをプロットしたものである。概していうと、（従来法）PE-IWLS で推定した D と REM の V が似ている。つまり、集団間変動の概略は、REM 法でも従来法でも捉えられている。ただし、V にだけ、残基 160-162 の高いピークの隣に、半分くらいの高さの残基 169-172 のピークがある。この位置は、それぞれ、DFG モチーフ、TGY モチーフである。TGY にピークを生じた原因は、PDB ID 3DS6 の 2 構造によるが、これは、使用したデータ中唯一の I-1/2 型阻害薬複合体の構造である。一般に、キナーゼ阻害薬には、ゲフィチニブのような DFG-in 構造を認識する I 型と、イマチニブのような DFG-out に結合する II 型がある。I-1/2 型は、DFG-in のヒンジ領域に結合するが、II 型ファーマコフォアの結合する領域にも結合する。このピークは構造クラスタリングの情報を利用したために発生しているといえる。これらの DFG と TGY のピークは、残基 30（本来は 34）あたりの高いピーク（リン酸基アンカー領域）と連動しているのがみてとれる。また、図には示していないが、集団内変動 S の第二および第三主軸では、第 1 集団の構造によるピークがみられた。これは PDB ID 2LGC であり、使用したデータ中唯一の NMR 構造である。これは、構造データを集団に分割したために顕わになったものと考えられた。このように、構造クラスタリングと REM 法の組み合わせにより、阻害薬の結合に重要な部位の同定や、特徴構造の同定を同時に行えた例を確認できた。

（５）まとめ

タンパク質構造のダイナミクスや異質性の解析において、共分散行列は、分散と相関係数に分解して、あるいは、固有値と固有ベクトルに分解して解釈されることが一般的である。数値実験の結果からは、それらの推定精度において、REM 法が TS 法に劣るケースはほどんどなく、むしろ、集団サイズが小さい場合には、REM 法の精度が高いケースを多く確認できた。TS-OLS は概して推定精度が低かった。構造重ね合わせに繁用されるのは（一段階型）OLS であるが、この等分散の仮定が構造重ね合わせのアーチファクトの一因として考えられる。一方、TS-IWLS は、mi=3 の場合は、稀に、収束しない場合があった。REM 法は計算時間を要するという欠点があるが、集団サイズが大きくない場合は、それも問題にならないと思われるため、精度の点から REM 法の使用が好ましいと思われた。

PDB に登録されている p38α の実際の結晶構造への応用においては、構造クラスタリングと併用することで、これまでに報告されている構造と機能の関係（リン酸基アンカー、DFG モチーフ、TGY モチーフ）と整合性の取れた集団間変動を求めることができた。また、同時に、集団間・集団内変動として、特徴的な構造個体を同定することもできた。REM 法は、このような集団サイズのあまり大きくないデータセットへの適用が特に期待できることが明らかとなり、今後、さらなる用途の開拓を継続したい。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 Takashi Amisaki
２．発表標題 Variance Estimation for Conformational Variability Using Random Effect Models
３．学会等名 2021年 日本バイオインフォマティクス学会年会（IIBMP2021）
４．発表年 2021年

１．発表者名 Takashi Amisaki
２．発表標題 A Method for Comparing Structural Ensembles: Applications to Molecular Dynamics Trajectory Data
３．学会等名 CBI学会2019年大会
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関