参考資料配布

PRESS RELEASE

2023 年 9 月 26 日
理化学研究所、奈良先端科学技術大学院大学、
自然科学研究機構　生命創成探究センター

大量の回折データから異なる構造情報を見いだす方法

ータンパク質の多様な構造決定を実現するためのガイドラインー

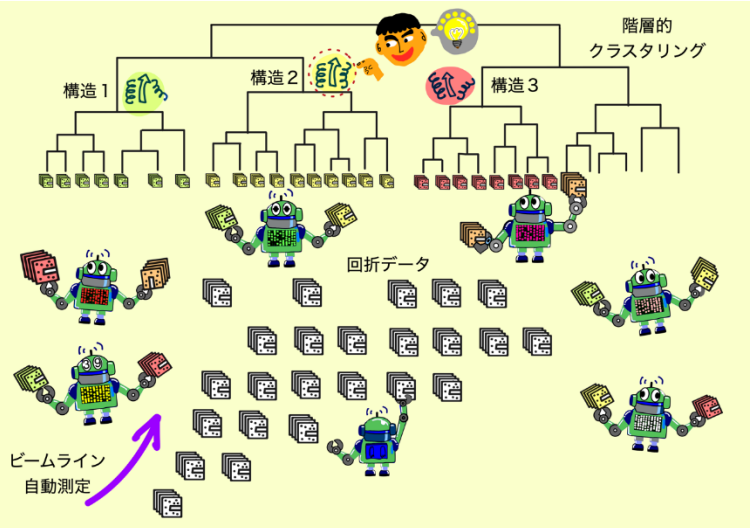
概要

理化学研究所（理研）放射光科学研究センター利用システム開発研究部門生物系ビームライン基盤グループ生命系放射光利用システム開発チームの平田邦生専任技師、山本雅貴グループディレクター、奈良先端科学技術大学院大学物質創成科学領域の藤間祥子准教授、自然科学研究機構生命創成探究センターの村木則文助教（研究当時）らの共同研究グループは、X 線結晶構造解析[1]において自動測定などによって得られた大量の回折データを分類し、タンパク質の多様な構造を捉えるための解析方法を提案し、広く一般に利用できるガイドラインを示しました。

本研究成果は、X 線を利用したタンパク質の立体構造解析に新たな価値をもたらし、生命現象の基礎的理解の促進に貢献するものと期待されます。

今回、共同研究グループは機械学習アルゴリズム[2]の一種である「階層的クラスタリング（HC）[3]」を利用し、多数のタンパク質結晶から得られた回折データの中から、異なる構造（構造多型）に由来するデータを分類するための最適条件を見つけました。この条件を自動測定で大量に得られたタンパク質結晶の回折データに適用した結果、異なる結晶において構造多型が見られることはもちろん、同一結晶内であっても構造多型が見られる場合があることを示しました。

本研究は、科学雑誌『*Acta Crystallographica Section D*』（9 月 25 日付）に掲載されました。



階層的クラスタリング（HC）による多型抽出のイメージ

背景

タンパク質は生命現象を担う基本要素の一つです。タンパク質の構造はその働きと密接に関係しています。タンパク質の構造を明らかにする主要な手段の一つがX線結晶構造解析です。この手法では、試料としてタンパク質の「結晶」を利用することから、得られる立体構造情報は時間的、空間的に平均化されたものとなります。そのため、実際のタンパク質が持つ柔軟性による多様な構造を捉えることは難しいと考えられています。

大型放射光施設「SPring-8」[4]は、高性能なX線を利用できる世界有数の施設です。現在、理研ターゲットタンパクビームライン BL32XU や生体高分子結晶解析 II ビームライン BL45XU では、「全自動データ収集システム ZOO[5]」を用いた効率的な回折データの収集が可能になっています[注1)]。無人の自動測定によって、多数の結晶から短時間で回折データを収集でき、事前に入念な選定をしなくても、良質な回折データの選択が可能です。

ただし、大量の回折データの中には、タンパク質の機能に関わる構造の違い（構造多型）が予期せず埋もれていることがあります。しかし、これまで大量に得られた回折データから構造多型を効率的に分類する方法はありませんでした。本研究では、機械学習アルゴリズムの一種である「階層的クラスタリング（HC）」によって、多くのタンパク質結晶から取得した大量の回折データを分類する方法を模索し、効率的に構造多型を見いだす条件を検討しました。

注1） 2019年2月7日プレスリリース「タンパク質結晶から自動でデータ収集する「ZOOシステム」を開発」
https://www.riken.jp/press/2019/20190207_1/

研究手法と成果

共同研究グループは、まずタンパク質結晶の試験データを利用して、HCが構造多型の分類に適しているかどうかを調べました。HCでは回折データ間の類似性に基づいてデータを分類するため、類似性の定義やデータの利用方法が成功の鍵となります。類似性の適切な定義や最適な利用方法を見つけるため、トリプシンというタンパク質の中に、化合物（ベンズアミジン[6]）が存在している「化合物あり」と存在していない「化合物なし」の結晶から収集した回折データから、それぞれ50セットずつ計100セットの仮想データを作成し、2種類のHCで分類しました。

本来、結晶構造解析では、回折データを解析した後に得られる電子密度図[7]からタンパク質や化合物の構造を解釈します。一方、HCに類似度の指標として回折強度の相関係数（CC）[8]（強度相関）を利用すると、電子密度図を求める前にタンパク質結晶中の化合物の有無を検出できることが分かりました（図1）。

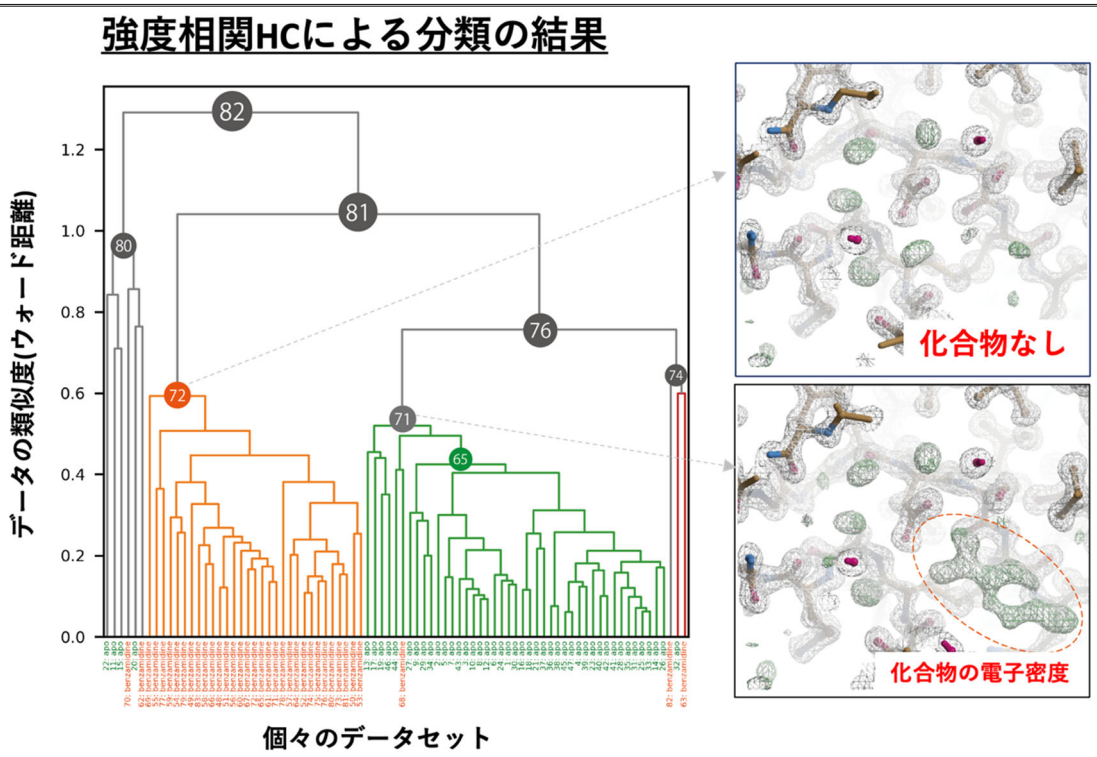


図 1 強度相関 HC に基づく 2 種類のデータの分類結果

トリプシンの中に化合物あり・化合物なしのデータを、回折強度の相関係数（強度相関）を用いて、階層的クラスタリング解析（HC）によって分類した結果。左図のデンドログラム（樹形図）が得られ、データの類似度によって個々のデータセットが分類される。この樹形図では、オレンジ色（化合物なし）と緑色（化合物あり）のデータが、番号 72 と 71 のところでまとまったグループを形成していることが分かる。該当する番号のデータから化合物の電子密度を計算すると（右図）、化合物の有無の違いは明らかである。

この試験結果と数値シミュレーションを通して、HC をうまく利用すると、多数のデータの中で 1%程度の構造の違い（相関係数が 99%）があれば、そのグループを検出して分類できることを突き止めました。さらに、HC の結果から、データに多型構造が含まれるかどうかを判定する指標（同型閾値[9]）も見つけました。

この同型閾値の有効性を、実際のタンパク質試料を用いた解析によって評価した結果、核内輸送受容体タンパク質ーNLS ペプチド複合体[10]のデータでは、異なる結晶において、二つの異なるペプチド結合様式（構造多型）が見いだされました。興味深いことに、同一結晶内でも構造の異なるデータが得られる場合もあることが判明しました（図 2）。

HypD タンパク質[11]でも、同様な解析を実施することで、タンパク質の内部・末端の形が全く異なる構造多型を、タンパク質の電子密度図を見る前に分類できました。さらに、この試料では、同じ結晶化母液から取り出した結晶間において、このような構造多型が見つかりました。

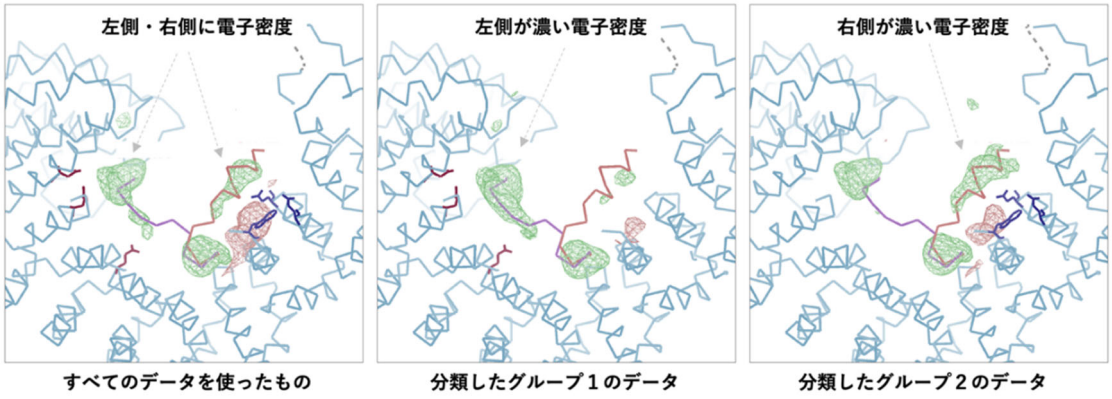


図２ タンパク質の構造多型解析の応用事例

核内輸送受容体タンパク質ーNLS ペプチド複合体のデータセットを分類しなかったもの（左：従来法）と、HC を適用し同型閾値を利用してデータを分類したもの（中央・右：今回の手法）の電子密度図（網目の部分）を比較している。各図の中央にある重要なペプチドの電子密度図に注目すると、データの分類により二つの構造多型を見いだすことに成功したことが分かる。

今後の期待

本研究では、大量に得られた回折データから HC と同型閾値を利用して、タンパク質の構造多型同士を効率的に分類する方法を提案し、広く一般に利用できるガイドラインを示しました。

本手法では、実際のタンパク質結晶試料において、同じ母液から得られた結晶や同一の結晶内からでも構造多型を検出することが可能です。また、タンパク質の電子密度図を見る前に大量の回折データを自動的に分類できるため、自動化による回折データ測定の効率化と併せて、多様な構造を高効率で捉える可能性が高まります。さらに、従来法では見失われていた構造がある場合や、複数の構造が混ざり合って不明瞭な場合においても、より確からしい解析が可能になります。

タンパク質のより確からしい構造や、機能を持つための変形過程にある多様な構造を捉えられるようになることから、生命現象の化学的理解が深まるだけでなく、創薬研究や触媒開発などの応用にもつながることが期待されます。

論文情報

<タイトル>
Elucidating polymorphs of crystal structures with intensity-based hierarchical clustering analysis on multiple diffraction datasets
<著者名>
Hiroaki Matsuura, Naoki Sakai, Sachiko Toma-Fukai, Norifumi Muraki, Koki Hayama, Hironari Kamikubo, Shigetoshi Aono, Yoshiaki Kawano, Masaki Yamamoto and Kunio Hirata
<雑誌>
Acta Crystallographica Section D Structural Biology

＜DOI＞
10.1107/S2059798323007039

補足説明

[1] X 線結晶構造解析
対象とする分子などの結晶を作製し、その結晶に X 線を照射して得られる回折データを解析することで、物質内部の原子の立体的な配置を調べる方法。この方法によって、タンパク質などの複雑な分子の立体構造を詳細に知ることができる。

[2] 機械学習アルゴリズム
アルゴリズムはある問題を解く手順を、単純な計算や操作の組み合わせとして定義したもの。機械学習は学習する手順のアルゴリズムを組み立てて、コンピューターにプログラミングして実行する。ここでは機械学習のために組み立てたアルゴリズムのことを指す。

[3] 階層的クラスタリング（HC）
教師なし機械学習に分類される、データ分類手法の一つ。X 線回折データの分類においても利用されてきた。正解ラベルやいくつのデータのグループが存在するかをあらかじめ仮定する必要はなく、データ同士の類似度を総当たり的に調べることで、類似しているデータと遠縁のデータを分類する。HC は Hierarchical Clustering の略。

[4] 大型放射光施設「SPring-8」
理研が所有する、兵庫県の播磨科学公園都市にある世界最高性能の放射光を生み出す施設。SPring-8 の名前は Super Photon ring-8 GeV に由来する。放射光（シンクロトロン放射）とは、電子を光とほぼ等しい速度まで加速し、電磁石によって進行方向を曲げたときに発生する細くて強力な電磁波のこと。SPring-8 では、遠赤外線から可視光線、軟 X 線を経て硬 X 線に至る幅広い波長域で放射光が得られるため、原子核の研究からナノテクノロジー、バイオテクノロジー、産業利用や科学捜査まで幅広い研究が行われている。

[5] 全自動データ収集システム ZOO
SPring-8 で開発した X 線結晶構造解析のための自動データ収集システム。実験結果を判断材料としてビームラインの装置群を制御し、実験者が事前に準備した測定条件表に従って、タンパク質結晶や低分子結晶からの無人自動測定を遂行できる。

[6] ベンズアミジン
タンパク質分解酵素であるトリプシンの阻害剤。X 線結晶構造解析において、標的タンパク質の分解を防ぐために用いられる。

[7] 電子密度図
回折強度データを解析することで得られる原子の位置を 3 次元の電子密度で表現した図。結晶構造解析では、電子密度図に原子モデルを当てはめて分子構造を解釈する。

[8] 相関係数（CC）

2 種類の測定値の直線的な関連の強さを表す指標。CC は Correlation Coefficient の略。

[9] 同型閾値

階層的クラスタリングによる分類の結果から、構造多型の有無や数を判別するために本論文で提示した閾値（しきいち）。

[10] 核内輸送受容体タンパク質ーNLS ペプチド複合体

真核生物では、核内で働くタンパク質（積荷タンパク質）は核ー細胞質間を自由に通過できず、核内輸送受容体タンパク質により輸送される。積荷タンパク質には、核内輸送受容体に識別されるタグ配列（NLS 配列）が存在する。核内輸送受容体タンパク質ーNLS ペプチド複合体は、核内輸送受容体タンパク質と NLS 配列のみのタンパク質の断片（ペプチド）が結合した構造を持つ。

[11] HypD タンパク質

ヒドロゲナーゼに必要な金属補因子を運搬するタンパク質の一つ。ヒドロゲナーゼは水素（H）をプロトン（H^+）に酸化する酵素であり、その酵素活性には HypD タンパク質が必須である。

共同研究グループ

理化学研究所　放射光科学研究センター　利用システム開発研究部門
生物系ビームライン基盤グループ
特別研究員　松浦滉明　（マツウラ・ヒロアキ）
研究員（研究当時）　坂井直樹　（サカイ・ナオキ）
（現　高輝度光科学研究センター　構造生物学推進室　研究員）
グループディレクター　山本雅貴　（ヤマモト・マサキ）
生命系放射光利用システム開発チーム
専任技師　平田邦生　（ヒラタ・クニオ）
専任技師　河野能顕　（カワノ・ヨシアキ）
奈良先端科学技術大学院大学　物質創成科学領域
准教授　藤間祥子　（トウマ・サチコ）
大学院生　端山浩輝　（ハヤマ・コウキ）
教授　上久保裕生　（カミクボ・ヒロナリ）
自然科学研究機構　生命創成探究センター
助教（研究当時）　村木則文　（ムラキ・ノリフミ）
（現　慶應義塾大学　理工学部　化学科　准教授）
教授　青野重利　（アオノ・シゲトシ）

研究支援

本研究は、日本学術振興会（JSPS）科学研究費助成事業新学術領域研究（研究領域提案型）「動的構造解析に資する固定ターゲット微小結晶解析法の開発（研究代表者：山本雅貴）」、同基盤研究（C）「人工知能を有する自動回折データ収集システムの開発（研究代表者：平田邦生）」、日本医療研究開発機構（AMED）生命科学・創薬研究支援

基盤事業（BINDS）「創薬等ライフサイエンス研究のための相関構造解析プラットフォームによる支援と高度化（SPring-8/SACLA におけるタンパク質立体構造解析の支援および高度化）（領域代表：山本雅貴）」「生命科学と創薬研究に向けた相関構造解析プラットフォームによる支援と高度化（領域代表：山本雅貴）」、科学技術振興機構（JST）研究成果展開事業研究成果最適展開支援プログラム（A-STEP）産学共同（本格型）「迅速微量多検体構造解析を可能とする無細胞タンパク質結晶化技術の開発（研究責任者：上野隆史）」の助成を受けて行われました。

発表者・機関窓口

＜発表者＞　※研究内容については発表者にお問い合わせください。
理化学研究所　放射光科学研究センター　利用システム開発研究部門
　生物系ビームライン基盤グループ
　　グループディレクター　　山本雅貴（ヤマモト・マサキ）
　　生命系放射光利用システム開発チーム
　　専任技師　　平田邦生（ヒラタ・クニオ）

奈良先端科学技術大学院大学　物質創成科学領域
　　准教授　　藤間祥子（トウマ・サチコ）

自然科学研究機構　生命創成探究センター
　　助教（研究当時）　　村木則文（ムラキ・ノリフミ）

＜機関窓口＞
理化学研究所　広報室　報道担当
Tel: 050-3495-0247
Email: ex-press [at] ml.riken.jp

奈良先端科学技術大学院大学　企画総務課　渉外企画係
Tel: 0743-72-5026/5063　Fax: 0743-72-5011
Email: s-kikaku [at] ad.naist.jp

自然科学研究機構　生命創成探究センター　研究戦略室
Email: press [at] excells.orion.ac.jp

※上記の[at]は@に置き換えてください。