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

平成　２９　年　　６　月　　７　日現在

機関番号：　１４４０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６４４００７８

研究課題名（和文）混合正規分布モデルを用いた低解像度密度マップ・原子モデルの重ね合わせ比較法の開発

研究課題名（英文）Comparison and superposition of low-resolution density maps and atomic models using Gaussian mixture model

研究代表者

川端　猛（KAWABATA, TAKESHI）

大阪大学・たんぱく質研究所・寄附研究部門准教授

研究者番号：６０３４３２７４

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：電子顕微鏡による３次元密度マップは生体高分子の構造に関する重要な情報を含んでいる。これらのデータを活用するため、密度マップや原子モデルを問い合わせとして、データベース内の類似したマップやモデルを検索するWEBサービス「Omokage検索」を開発した。検索は、3D代表点の距離を変換した特徴量を用いて行う。マップやモデルの立体重ね合わせもWEBで可能であり、この重ね合わせには混合正規分布モデル(GMM)による形状近似表現を用いた。GMMへの変換を安定かつ高速に行うため、「ガウス関数入力型GMM」を開発した。複数の原子モデルを密度マップに重ねる方法やマップ内のヘリックスを認識する方法の開発も進めた。

研究成果の概要（英文）：3D density maps by electron microscopy have important information about biomolecular structures. To make a full use of these maps, we developed a WEB server “Omokage search” to search the global shape similarity of biological macromolecules in databases for 3D density maps and atomic models. The server searches using the distance-based profiles from the 3D points of the maps and models. It also performs 3D superimpositions using the Gaussian mixture model (GMM), for approximating shapes of the maps and the models. The new algorithm “Gaussian-input GMM” was developed for converting into GMMs more robustly and rapidly. We started to develop the methods for multiple subunit fitting and alpha-helix detection.

研究分野：　構造バイオインフォマティクス

キーワード：　構造・機能予測　電子顕微鏡　単粒子解析

１．研究開始当初の背景

生体高分子の立体構造データを得る方法として、X 線結晶構造解析と核磁気共鳴法の二つが最も広く使われ、分子機能に関する重要な知見をもたらしている。近年、これら二つを補完する第三の手法として、電子顕微鏡画像の単粒子解析(single particle analysis）による３次元密度マップが提供されはじめた。この方法では、サンプルの可溶化・結晶化が不要であり、ウィルスやリボゾームなどの巨大複合体の構造決定に向いており、他の手法では得られない重要な構造情報が得られる。３次元密度マップのデータは、EMDB というデータベースに集積され、現在(2013 年)、約 2000 エントリーが登録されている。一方、問題点も多い。その一つは、解像度は 10～30 Åとあまり高くないことである。これはドメインか二次構造がかろうじて識別できる程度の情報量であり、先験的に残基・原子レベルのモデルを構築することは大変難しく、サブユニットの位置さえ不明な場合も多い。そのため、電顕が専門でない研究者は、電顕の３次元密度マップを「どこがどれだかよくわからないぼんやりしたデータ」としてしか認識していない場合が多く、電顕でしか得られない貴重な情報を活用できていない。この低解像度の問題を解消するため、サブユニットの単体の構造が X 線結晶解析で解かれている場合は、低解像度の密度マップにサブユニットの原子モデルを重ね合わせて、複合体の原子モデルを構築する試みが行われている。うまく重ね合わせができた場合は、複合体の原子モデルを生成することができる。通常、この重ね合わせは分子表示ソフト(UCSF Chimera)を用いてマニュアルで行うことが多い。しかし、特にサブユニットの数が多数ある場合は、その重ね合わせをマニュアルで構築することは容易ではない。研究代表者は、密度マップと原子モデルを効率的に重ね合わせるために、統計学の分野で開発された混合正規分布モデル(Gaussian Mixture Model;GMM)を用いた分子表現法を 2008 年に提案した(Kawabata (2008) Biophys.J., 95, 4643)。GMM とは複数の正規分布の重み付き線形和のことである。上図のように、原子モデル、密度マップとも～100 個程度の正規分布で分子の概形は表現できる。通常、密度マップと原子モデルの重なりの度合いを計算するには、[マップの格子点の数(503～2563)] ×[原子モデルの原子数(103～104)] に比例する計算時間がかかるが、双方を GMM に変換すると、その計算時間は[正規分布の数]×[正規分布の数]にまで削減でき、劇的に計算速度が向上する。この 2008 年の研究では GMM の間の重なり、反発のエネルギーを導入し、ランダム探索と最急降下法の組み合わせで最適な配置を得る方法を採用し、そのプログラムは ***gmfit*** と命名された。

２．研究の目的

近年、電子顕微鏡単粒子解析から生体高分子複合体の低解像３次元密度マップを得る技術が進展し、多数のデータが集積している。こうした密度マップデータを有効活用するには、複数の密度マップ群の相互比較、および、サブユニットの原子モデルを密度マップに重ね合わせて複合体原子モデルを構築する計算が不可欠である。本研究では、これらの計算を効率的に行うことを目指し、密度マップや原子モデルを近似表現する混合正規分布モデル(Gaussian Mixture Model; GMM)を利用した計算手法の発展的な開発を行う。密度マップデータベースに対する検索、二つのマップ間の変形と差分の検出、一つのマップに複数の原子モデルを効率的に重ねる方法、重ね合わせの詳細な修正を行う機能などを開発する。これらの技術をもとに密度マップデータを有効活用できる計算機環境を構築する。

３．研究の方法

(1)3D 密度マップ・原子モデルデータベースに対する形状検索サービスの開発

EMDB や PDB にはリボゾーム、プロテアソームなど似た複合体の密度マップや原子モデルが多数登録されている。こうした類似構造群の形状を直接比較して検索できる WEB サーバの開発が望まれる。川端は GMM を用いたペアワイズの重ね合わせ計算プログラム ***gmfit*** を既に開発している。この重ね合わせは１ペアでは極めて高速だが、WEB サービスとして実装するには遅い。そこで、連携研究者の鈴木が開発した距離プロファイル法と組み合わせることを計画している。この方法は重ね合わせ操作をせずに類似性を計算できるため、計算は極めて高速である。

(2) ガウス関数入力型混合正規分布モデルのアルゴリズム開発

密度マップや原子モデルを GMM に変換する場合、入力点群に対する尤度を最大にする EM アルゴリズムを用いるのが一般的である。しかし、この方法では、マップの格子幅や原子半径が無視され、本来の大きさが再現されないこと、３個以下の入力点に対応するガウス関数が生じたときに、ゼロ除算のエラーが生じ異常終了する「特異性」の問題があること、格子数の多い密度マップの変換に計算時間がかかることなどの問題がある。これらの問題を解決するために、ガウス関数入力型の GMM のアルゴリズムの開発を行う。

(3) 複数サブユニットのフィッティングのためのアルゴリズムの開発

複数のサブユニットを一つの密度マップに重ねる計算は、計算量・モデル精度の両方において、困難な問題である。本研究では、複数のサブユニットの配置を効率的に探索する「セグメンテーション＆フィッティング」のアルゴリズムを提案し、さらに、部分的な実験情報を取り込むことで、より妥当な重ね

合わせ解を探索する。

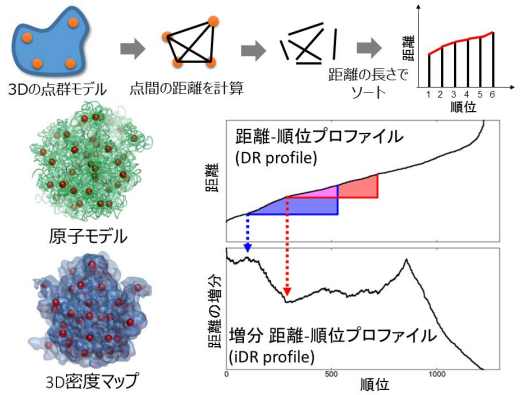
(4) 高解像度マップからのデノボモデリングのための手法開発

2013 年頃から、4Å以下の高い解像度の密度マップが次々と報告されるようになり、既知の X 線構造を重ね合わせるだけではなく、密度マップだけから原子モデルを構築することが現実的となってきた。本研究では、その第一歩として、密度マップからαヘリックスを認識する手法の開発に着手する。

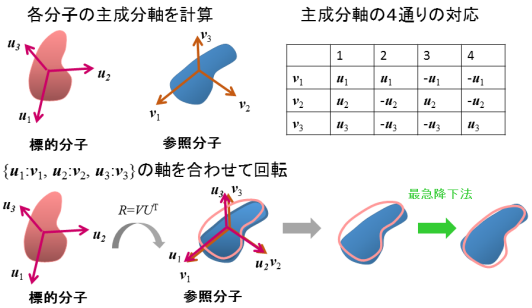
４．研究成果

(1)3D 密度マップ・原子モデル群に対する形状検索サービス「Omokage 検索」の開発

入力された問い合わせの密度マップや原子モデルと類似した形状のデータを検索する WEB サービス「***Omokage*** 検索」を開発した。検索対象は EMDB 内の 3 次元密度マップ群及び PDB 内の原子モデル群の両方である。密度マップ、原子モデルの両方とも、あらかじめ３次元の点群および GMM に変換され、サーバに収納されている。サーバは、まず、点群間の距離から作られた１次元の特徴ベクトル (iDR-profile; incremental distance rank profile; 増分 距離-順位プロファイル)を検索する。検索はおおむね１分以下で終了する。



検索された類似構造と問い合わせ構造を重ね合わせて表示することもできる。この重ね合わせは、プログラム ***gmfit*** の「１対１」の重ね合わせ機能を用いて実装している。三つの主成分軸を合わせた初期構造をもとに、GMM で最急降下法で姿勢を最適化する。

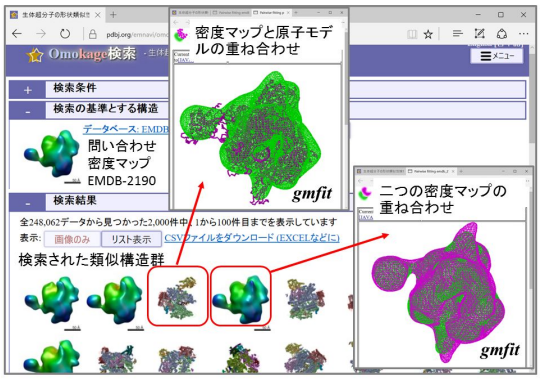


	1	2	3	4
v_1	u_1	u_1	$-u_1$	$-u_1$
v_2	u_2	$-u_2$	u_2	$-u_2$
v_3	u_3	$-u_3$	$-u_3$	u_3

検索・重ね合わせは、密度マップ・原子モデルの区別なく行うことができる。また、本手法は、複合体全体の大域的な比較であるため、円順列変異(circular permutation)や分子擬態(molecular mimicry))など、構成する鎖の数、配置、種類が大きく異なるような類似性でも認識できる。

提案手法の性能評価として、既存の形状比較サーバ ***EMSURFER*** (Esquivel-Rodríguez, J. et al., *BMC Bioinformatics* (2015), 16:181) のスコアと本手法のスコアを、ClpB-ClpP と 70S ribosome のデータセットを用いて比較した。その結果、本手法のほうが、解像度や体積が異なる類似分子の認識に優れていることがわかった。

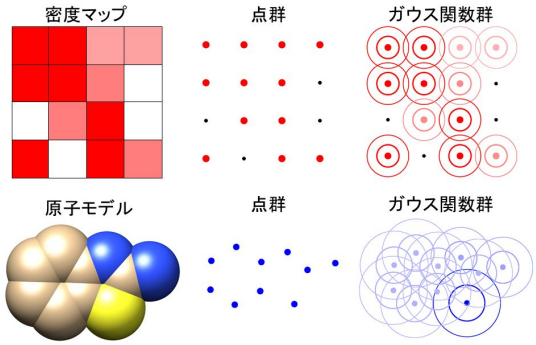
このサーバについては、Bioinformatics 誌、Nucleic Acids Research 誌に報告した(②Suzuki et al., 2016; ①Kinjo et al., 2017)。



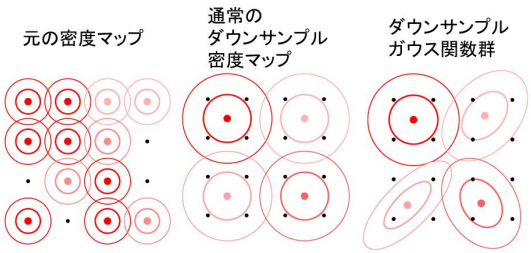
論文発表後も、ユーザーが密度マップや原子モデルをアップロードする機能を追加、X 線小角散乱によるダミー原子モデルも取り込むなどの改良を加えている。

(2)ガウス関数入力型混合正規分布モデルのアルゴリズム開発

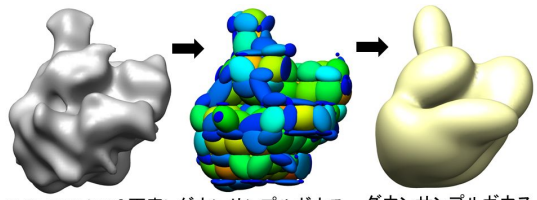
密度マップや原子モデルを GMM に変換する過程は、重ね合わせ計算に先立って必要な重要な過程である。これまで、この計算は点群を入力とした尤度関数を EM アルゴリズムで解く標準的な算法を採用してきた。しかし、この方法では、入力データの大きさが無視されること、特異性の問題で異常終了する場合があること、大規模な密度マップに対する計算が遅いことの三つの問題があった。これらの問題を解決するため、ガウス関数群を入力とする GMM を考え、その尤度関数を EM アルゴリズムで解く算法を考案した。入力をガウス関数群とすると、尤度関数を解析的に求められなくなるが、Q 関数の最適化は可能であるため、通常の点群の最適化と同様の手続きで、尤度を最大化することができる。この算法では、定式上、GMM を構成するガウス関数の分散が入力ガウス関数の分散を下回ることはないため、特異性の問題は生じない。実際に多数のマップについて GMM への変換を試みると、点群入力の GMM に比べ、より正確な慣性半径の GMM に異常終了を全く起こさずに変換することができることがわかった。



また、大規模な密度マップを高速に GMM に変換する手法についても開発を行った。通常、画素数の多い密度マップの画像処理を高速に行いたい場合、いくつかの近傍の画素群(2^3,3^3,4^3,…個)を加算して一つの画素にした、画素数の少ないダウンサンプルマップを作成し、それに処理を施すことが多い。本研究では、近傍の画素を一つの非等方的ガウス関数（近傍点群の重み付き平均と共分散を持つ）として融合し、多数のガウス関数群を入力として、前述のガウス関数入力型の GMM を適用する方法、「ダウンサンプルガウス関数法」を提案した。



この方法では、元のマップにより近い GMM を、高速に生成することができる。また、生成された GMM は、元のマップと同じ慣性半径や共分散を持つという特長がある。



EMD-1282 (120^3 画素). Human RNA polymerase II.　ダウンサンプルガウス関数群. 8^3画素を1つのガウス関数に統合.　ダウンサンプルガウス関数群を入力として作成したGMM. 10個のガウス関数を使用.

このプログラム ***gmconvert*** のソースはホームページ③で公開されている。また、この成果は、2017 年 5 月末に雑誌投稿を行い、現在査読中である。

(3) 複数サブユニットのフィッティングのためのアルゴリズムの開発

複数のサブユニットを一つの密度マップに重ね合わせる「多対１」の重ね合わせを効率的に行うために、まず「セグメンテーション＆フィッティング」法と呼ぶ算法を開発した。この方法は、まず、サブユニットの初期配置を生成し、配置に従い密度マップを各サブユニットの領域に分割（セグメンテーション）する。次に、各サブユニットを分割された領域だけ考慮して重ね合わせる（フィッティング）。この手続きを収束するまで繰り返すという算法である。この方法はこれまでの方法より効率よく、全マップを覆うような配置を求めることができる。しかし、低解像度マップが対象の場合、この算法だけでは正しい配置が得られない場合も多い。その場合、部分的な実験情報を取り込むことが重要と考え、対称性やサブユニット間の近接性を取り込めるように探索法やスコア関数を工夫した。複数サブユニットの重ね合わせ問題は、本質的に困難な問題であり、現在も開発を続けている段階である。この研究の学会発表は既に行っている(⑦、⑩)。内容がまとまり次第、論文をまとめ投稿したい。

(4) 高解像度マップからのデノボモデリングのための手法開発

デノボモデリングの手始めとして、αヘリックスを認識するための GMM の算法の開発を行った。通常の GMM では、推定されるガウス関数の大きさや形状に制限はない。そこで、様々な長さのαヘリックスの形状に対応するガウス関数群のライブラリを用意し、その中から選ぶという制限を設けた「ライブラリーGMM」の算法を考案した。試験計算の結果、この算法である程度ヘリックスの位置を同定できることがわかった。しかし、より正確にαヘリックスの位置、長さ、向きを決定するには、ヘリックスを一つのガウス関数で表現するモデルは不十分であり、より原子モデルの詳細を持ったモデルで、精密化を行う必要があることも明らかになった。実際にアミノ酸配列に対応する原子モデルを構築するには、各部位のアミノ酸を決定する必要があり、側鎖構造の構築も必要となる。デノボモデリングは、複雑で裾野の広い分野であり、一つの算法で全てが解決するわけではなく、二次構造要素識別、配列からの二次構造予測、側鎖生成、ループ接続、精密化など、様々な要素技術を統合して用いる必要がある。「ライブラリーGMM」は、少なくとも、αヘリックスの候補群を算出するには有効であることは確かなので、他のプログラムとの比較等を進め、できるだけ早く論文を執筆する予定である。学会発表は①で行っている。

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 Kinjo, A.R., Bekker, G.J., Suzuki, H., Tsuchiya, Y., Kawabata, T., Ikegawa, Y., Nakamura, H. Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Res.*, 査読有, Vol. 45(D1), 2017 年, D282-D288.

② Suzuki,H., Kawabata,T., Nakamura,H. Omokage search: shape similarity search

service for biomolecular structures in both the PDB and EMDB. *Bioinformatics*, 査読有,Vol. 32、2016 年, 619-620.

〔学会発表〕(計 12 件)

① 川端 猛, 中村春木,電顕３次元密度マップからαヘリックスを認識する混合正規分布モデルの開発,日本生物物理学会第 54 回年会, 2016.11.27, つくば国際会議場.

② 鈴木博文, 川端 猛, Gert-Jan Bekker, 中村春木, EMDB, PDB, SASBDB 中の多階層構造データを対象とするウェブベースのサービス, 日本生物物理学会第 54 回年会, 2016.11.27, つくば国際会議場.

③ Suzuki, H., Kawabata, T., Nakamura, H., Omokage search: shape similarity search for PDB atomics models, cryo-EM map data, and SAXS dummy atom models, 第 16 回日本蛋白質科学会年会, 2016.6.9, 福岡国際会議場.

④ Kawabata, T., Suzuki, H., Nakamura, H., Omokage search and gmfit : shape similarity search and superposition among models and maps, Biophysical Society 60-th Annual Meeting, 2016.2.28, Los Angeles.

⑤ Kawabata, T., Omokage search and gmfit : shape similarity search and superposition among models and maps. The wwPDB Foundation 主催国際シンポジウム "Integrative Structural Biology with Hybrid Methods", 2015.10.3, 大阪大学会館.

⑥ 鈴木博文, 川端 猛, 中村春木, EMDB と PDB データの形状類似検索:Omokage 検索,日本生物物理学会第 53 回年会, 2015.9.13, 金沢大学角間キャンパス.

⑦ 川端 猛, 鈴木博文, 中村春木, 低解像度密度マップへの複数のサブユニットのあてはめ計算 – 実験情報による拘束の利用 -, 日本生物物理学会第 53 回年会, 2015.9.15, 金沢大学角間キャンパス.

⑧ Suzuki, H., Kawabata, T., Nakamura, H., Shape similarity search of EMDB and PDB: Omokage search, 第１５回日本蛋白質科学会年会, 2015.6.26, あわぎんホール（徳島市）.

⑨ Kawabata, T., Suzuki, H., Nakamura, H., Omokage search : A rapid shape similarity search and superposition of biological assemblies, Biophysical Society 59-th Annual Meeting, 2015.2.19, Baltimore.

⑩ 川端 猛, 鈴木博文, 中村春木, セグメンテーション & フィッティング - 低解像度密度マップへの複数のサブユニットのあてはめ計算法 –, 日本生物物理学会第52回年会, 2014.9.27, 札幌コンベンションセンター.

⑪ 川端 猛, 鈴木博文, 中村春木, 構造データベース中の３次元電子顕微鏡データの形状比較とフィッティング, 日本生物物理学会第52回年会, 2014.9.27, 札幌コンベンションセンター.

⑫ 鈴木博文, 中村春木, 構造データベースの形状類似検索, 第 14 回日本蛋白質科学会年会, 2014.6.27, ワークピア横浜/横浜産貿ホールマリネリア.

〔その他〕

ホームページ等

① ***Omokage* 検索：**
https://pdbj.org/emnavi/omo-search.php

② ***Pairwise gmfit* のサービス：**
https://pdbj.org/gmfit/

③ ***gmfit* /*gmconvert* のプログラム配布**
http://strcomp.protein.osaka-u.ac.jp/gmfit/

６．研究組織

(1)研究代表者

川端 猛 （KAWABATA TAKESHI） 大阪大学・蛋白質研究所・寄附研究部門准教授

研究者番号：60343274

(3)連携研究者

鈴木 博文（SUZUKI HIROFUMI）大阪大学・蛋白質研究所・特任助教

研究者番号：60418572