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　３ 年　 ５ 月 ２１ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ０７３６４

研究課題名（和文）混合正規分布を用いた電顕密度マップからのマルチスケール原子モデル構築法の開発

研究課題名（英文）Multiscale atomic modeling based on electron microscopy 3D map using Gaussian mixture model

研究代表者

川端　猛（Kawabata, Takeshi）

大阪大学・蛋白質研究所・特任准教授（常勤）

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）：低温電子顕微鏡による近原子レベルの解像度の３次元密度マップを基にした、様々な原子モデルの構築手法を、混合正規分布モデル(GMM)を用いた分子表現を利用して開発した。まず、分子、マップをその大きさが保存されるように変換できる「ガウス関数入力型GMM法」、「ダウンサンプルガウス関数法」を開発した。また、巨大複合体の空洞の同定に有効な「空洞ポケット（cave pocket)」を同定する算法もモルフォロジーを使って開発した。また、複数サブユニットの原子モデルをマップの一部に重ね合わせる場合に有効な「マスク付きセグメンテーション&フィッティング法」の開発も行った。

研究成果の学術的意義や社会的意義
電子顕微鏡による生体高分子の立体構造解析はここ数年で学術研究機関や企業研究者に広く普及し、密度マップから原子モデルを客観的に構築できるソフトウエアの確立が望まれている。本研究で開発した原子モデルのフィッティング法などの手法は、こうした原子モデルの構築に有用であるはずである。本研究成果の一つであるプログラムのソースはWebページで公開され、無償で使用できるように配慮されており、この分野の研究者に広く貢献することが期待できる。

研究成果の概要（英文）：We have developed several atomic modeling methods based on Cryo-EM 3D density map with near-atomic resolution using Gaussian mixture model (GMM). In order to convert to a map (or an atomic model) into a GMM with the same size, “Gaussian-input GMM” and “down-sampling GMM” methods have been invented. We also developed a new algorithm “cave pocket” to detect an internal hollow space in a large macromolecular complex using mathematical morphology. The method “masked segmentation & fitting” has been developed to fit multiple atomic models of subunits into a local region of 3D map.

研究分野：構造バイオインフォマティクス

キーワード：電子顕微鏡　単粒子解析　混合正規分布モデル　フィッティング　原子モデリング　EMアルゴリズム　モルフォロジー

１．研究開始当初の背景

X 線結晶解析、NMR に並ぶ第３の構造生物学として、低温電子顕微鏡が注目を集め始めていた。それまで、密度マップは、10-20Å程度の解像度のサブユニットやドメインの構造が判別できる程度のマップが多く、X 線結晶解析で解かれたサブユニットの原子モデルを重ねることで原子モデルを構築してきた。2013 年に、Yifan Cheng のグループは 3.275Åの密度マップを報告した。このマップからはヘリックスのピッチや大きなアミノ酸の側鎖を識別できる。このレベルの解像度は近原子レベルの解像度と呼ばれ、この後、3Å前後の解像度の密度マップが相次いで一流誌に報告されるようになった。これにより、密度マップからの原子モデルの構築法が大きく変化する可能性がでてきた。X 線結晶構造が全く無くても、5Å程度の解像度があれば二次構造要素を構築でき、3Å前後のマップからは、残基レベル、原子レベルの立体構造を最初から構築できる可能性がある。しかし、標準的な X 線結晶解析の解像度（1.5-2.5Å）に比べると劣るため、解析プログラムで半自動的に原子モデルを構築するのは難しい。GUI ツールを使って専門家が手作業で初期モデルを作り、それを精密化するという方法で構築する場合が多い。このプロセスの計算機プログラムを用いた効率化、客観化が望まれていると考えられた。申請者は、これまで混合正規分布モデル(Gaussian Mixture Model;　GMM)を用いて、10-20Åほどの解像度の電顕密度マップと原子モデルを重ね合わせる方法を開発し、プログラム ***gmfit***、WEB サービス ***Omokage*** 検索を開発してきた。混合正規分布モデルとは、密度マップをガウス関数の線形和で表す手法である。密度マップ（多数の格子点群）、原子モデル（多数の球群）を、1〜40 個ほどの３次元の異方的ガウス関数の和に変換する。ガウス関数の場合、この変換を効率的に行う算法(EM アルゴリズム)が知られている。GMM 間の重なりを最大化する配置は、重なりの勾配などを用いて求める。ガウス関数どうしの重なり積分は解析的に求められるため、重なり最大化の探索も高速に行うことができる。本研究では、この GMM を用いた手法を発展させ、高解像度化が進む電顕の密度マップから、解像度に応じた様々なレベルの詳細さ（マルチスケール）で、モデルを構築する算法・プログラム群を開発することを目的とする。。

２．研究の目的

低温電子顕微鏡の技術革新により、3〜4Åの近原子レベルの解像度の３次元密度マップが報告されるようになった。10〜20Åのマップではサブユニットの原子モデルの重ね合わせしかできなかったが、高解像度の密度マップからは、事前の知識なしにマップだけから原子モデルを構築できる可能性がある。本研究では、様々な解像度の電顕３次元密度マップから、解像度に応じた詳細さで原子モデルを構築するための基本的な算法・プログラムを開発することを目的とする。このために、申請者がこれまで開発してきた混合正規分布モデルを発展させ、サブユニットだけではなく、二次構造要素（αヘリックスなど）やアミノ酸をガウス関数で表現し、これらマルチスケールな要素を密度マップに重ね合わせ、適合させることで原子モデルを構築する手法を開発する。　様々なレベルとは、サブユニット・ドメイン（10Å以上)、二次構造要素（5-10Å)、アミノ酸(3-5Å)、原子(3Å以下)の４つの段階を想定している。これらのモデルは等方的(球状)か異方的（楕円球状）なガウス関数で統一的に記述される。また、密度マップは低解像度の場合は上図のような GMM で表現されるが、高解像度では格子点ごとに等方的ガウス関数を置いたモデルで表現する。モデリングは、サブユニットの原子モデルが入手可能な場合、それを密度マップに重ね合わせる段階(前述のプログラム ***gmfit*** を改良して使用)から始める。次は、解像度に応じて、残ったマップの部分に原子モデルを構築していき、まず二次構造要素を同定するステップ、次に側鎖を生成させアミノ酸配列をあてはめるステップと続く。

３．研究の方法

以上述べた研究計画に従って、実際に行った研究の概略を以下に順に説明する。

(1)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

分子や密度マップを混合正規分布モデルへ変換は、本研究の基幹となる技術である。まず、入力する分子やマップの大きさが維持されるような変換のアルゴリズムの開発を行う。

(2)高分子複合体の「空洞ポケット」を計算する手法の開発

申請者はこれまで、タンパク質表面の結合ポケットを同定するプログラム ghecom を開発してきた。しかし、この手法では電顕で解析されるような巨大複合体の内部の空洞を同定するには不向きである。そこで、こうした空洞を同定するアルゴリズムの開発を行う。

(3)密度マップから二次構造要素を同定するプログラムの開発

密度マップから、αヘリックスやβシートなどの二次構造要素を同定することは、中程度のマップからのモデリングの基本的なステップとなるはずである。そこで、細長いガウス関数一つでαヘリックスを表現することで、効率的に二次構造要素を同定するプログラムを開発する。

(4)密度マップに複数の剛体サブユニットを重ね合わせるプログラム ***gmfit*** の機能拡張

複数個のサブユニットの原子モデルを、密度マップに重ね合わせる問題には、申請者は以前から取り組んでおり、***gmfit*** というプログラムを開発してきた。しかし、この問題は、サブユニットの数が多い場合、また、マップの一部に局所的にしかフィットしない場合は、探索が著しく難しくなる。そこで、多数サブユニットの局所フィッティングの場合でも、破綻しないようなアルゴ

リズムの開発を行う。

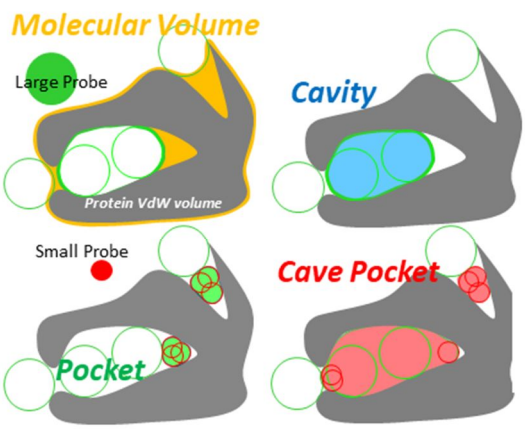
４．研究成果

(1)ガウス関数入力型混合正規分布モデルの開発

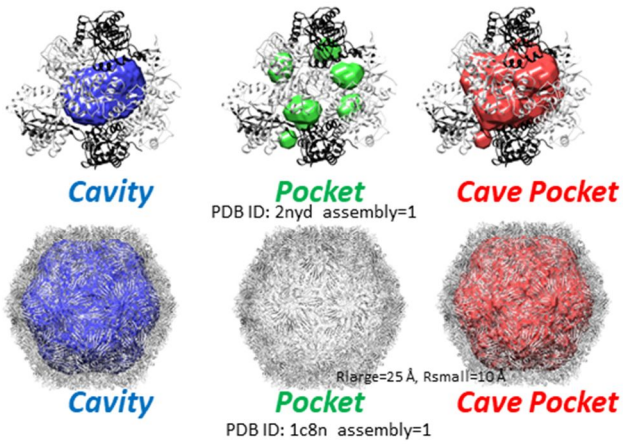
混合正規分布モデル(GMM)は、原子モデルや密度マップを近似表現するために用いられ、EM アルゴリズムで計算される。しかし、従来の手法では、入力となる原子群やボクセル群を点群として扱うため、原子半径やグリッド幅などの大きさを考慮できなかった。また、ガウス関数の分散がゼロに近くなると異常終了する特異性問題もある。これらの問題を改良するため、入力を点群ではなく、ガウス関数群とした GMM(Gaussian-input GMM)を EM アルゴリズムで解く方法（ガウス関数入力型 GMM）を新しく提案した(Kawabata, 2018)。本手法では原子やボクセルの大きさも考慮されるため、元となる入力マップや原子モデルの大きさ（分散と共分散行列）が再現され、特異性問題も解決できる。また、大規模な密度マップを入力とする場合は、近隣のボクセル群を非等方性ガウス関数に融合する、「ダウンサンプルガウス関数」という高速算法も提案、これを入力として EM アルゴリズムとして、「ダウンサンプル入力型 GMM」 も提案した。これらは、大規模な密度マップを GMM に高速に変換するために有効である。本研究は 2018 年に Journal of Structural Biology 誌にて出版されている。この算法はプログラム ***gmconvert*** に実装され、***gmfit*** の WEB サイトからダウンロード可能である。

(2)高分子複合体の「空洞ポケット」を計算する手法の開発

タンパク質の低分子化合物の結合部位は、「ポケット」形状をしていることが知られている。Kawabata & Go (2007)は、大小二つのプローブ球を用いて、「小さな球は入れるが、大きな球は入れない空間」がポケットであると定義した。その後 Kawabata(2010)は、ポケットを計算するプログラム ***ghecom*** を開発した。近年、シャペロニン、プロテアソームなど、内部に大きな空間を持つ巨大な高分子複合体の構造が多く決定されている。内部の空間が非常に大きい場合、閉鎖空間であっても、ポケットとして認識されない。そこで、Manak(2019)の提案に従い、「小さな球は入れるが、大きな球は外から入れない空間」という定義を導入し、「空洞ポケット(cave pocket)」と名付けた。別途、「空洞(cavity)」を、「プローブ球は入れるが、外へ出られない空間」と定義した。モルフォロジーでこれらの空間を厳密に定義し、その計算機能をプログラム GHECOM に組み込んだ。PDB 内の多くの分子について試験計算を行ったところ、空洞ポケットは、ポケットに比べて大きな内部空間の認識に有効であること、空洞に比べて分子内面の詳細な形状の検出に有効であることがわかった。本手法は、巨大な空洞をもつ高分子複合体の分子機能や安定性を理解する上で、有効なツールとなるだろう。本研究は、2019 年に Biophysic and Physicobiogy 誌にて出版されている。この算法はプログラム ***ghecom*** に実装されている。



プローブ球を用いた様々な空間領域の定義

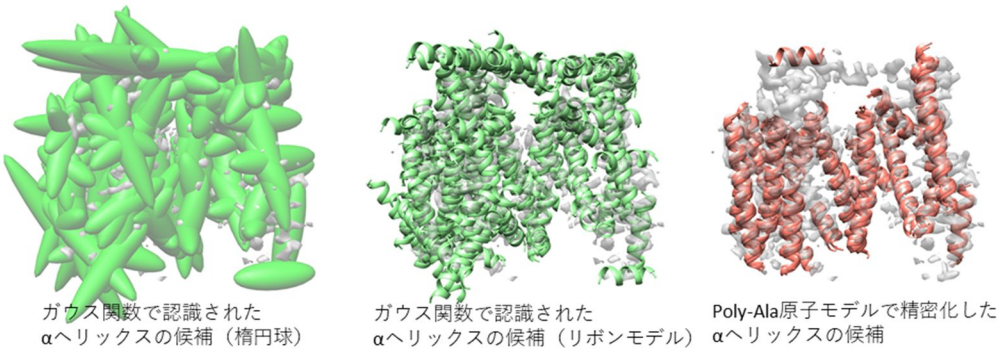


タンパク質複合体で同定された3種の空間領域

(3)密度マップから二次構造要素を同定するプログラムの開発

密度マップから、二次構造要素を同定することは、中程度のマップからのモデリングの重要なステップである。そこで、まず、細長いガウス関数一つでαヘリックスを表現することで、効率的に二次構造要素を同定するプログラムの開発を試みた。当初は、様々な長さのαヘリックスに対応するガウス関数のライブラリから選択する「ライブラリ GMM」という最尤法の枠組みで実装を始めたが、この枠組みではαヘリックスの本数が事前に既知であることが条件となってしまう。そこで、次のような２ステップで認識を行う方法を考案した。i) １ヘリックス－１ガウス関数表現で、できるだけ多数の候補ヘリックスを生成する。このとき、ガウス関数はラプラシアンをかけた形式(LoG: Laplacian of Gaussian)で行い、ランダムに配置したあと、最急降下法で姿勢の修正を行う。ii)前ステップで認識された候補ヘリックスについて、ポリアラニンで構成されたヘリックスのモデルに置き換え、最急降下法で姿勢の修正を行う。いくつかの電顕のマップにつ

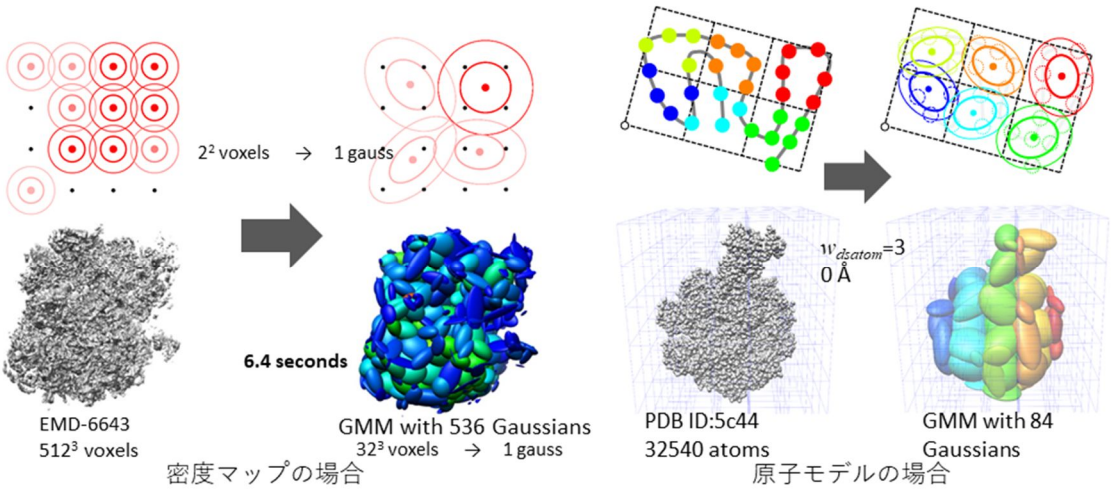
いて計算を行ったところ、良好な結果を得ている。今後は、さらなる精度と計算速度の向上、βシートなど他の二次構造要素への拡張を試みていきたい。



ガウシアン二次構造同定法の計算例. γ-secretaseの膜貫通部の３D密度マップ(EMD-3061)の場合.

(4)密度マップに複数の剛体サブユニットを重ね合わせるプログラム *gmfit* の機能拡張

複数個のサブユニットの原子モデルを、密度マップに重ね合わせる問題には、申請者は以前から取り組んでおり、***gmfit*** というプログラムを開発してきた。しかし、この問題は、サブユニットの数が多い場合、また、マップの一部に局所的にしかフィットしない場合は、探索が著しく難しくなる。そこで、多数サブユニットの局所フィッティングの場合でも、破綻しないようなアルゴリズムの開発を行う。一般に、マップ全体へのフィッティングに比べ、マップの一部へのローカルなフィッティングには、より詳細な形状表現が必要となる。混合正規分布モデルの場合、詳細な形状表現は、より多数のガウス関数を用いる必要があるが、多数のガウス関数の位置を推定するために、そもそも大きな計算量が必要となってしまう。そこで、密度マップの GMM の推定には、近隣のボクセル群を非等方性ガウス関数に融合する、「ダウンサンプルガウス関数変換法」



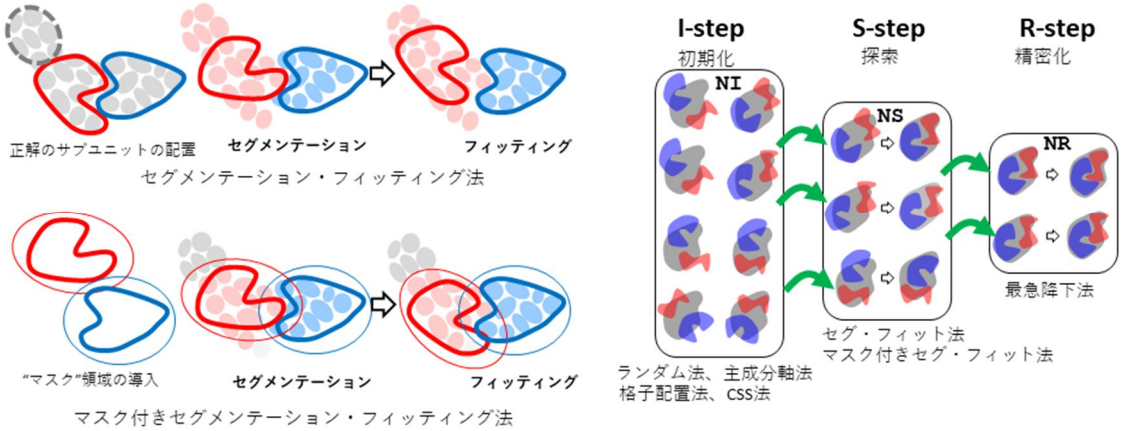
ダウンサンプルガウス関数変換法

という算法を提案した。さらに、原子モデルの GMM の推定についても、同様なアイデアで空間を格子に分割して、近隣の原子群を非等方性ガウス関数に融合する手法を開発した。これにより数百のガウス関数からなる GMM を非常に小さい計算時間で推定できるようになった。

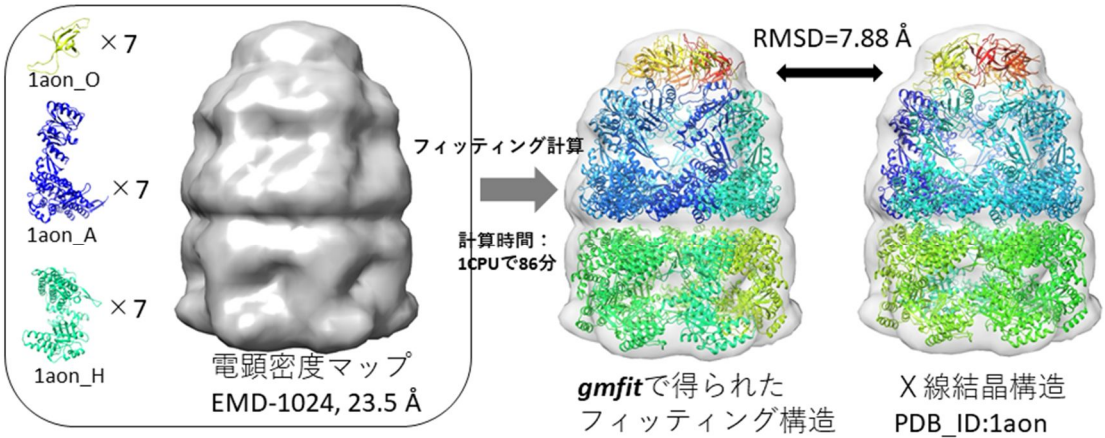
また、複数のサブユニットの原子モデルを一つのマップに重ねる手法として、まず「セグメンテーション・フィティング法」（セグ・フィット法）を開発した。これは、マップのサブユニットへの領域分割のステップ(segmentation)と、分割されたマップ領域だけにフィティングさせるステップ(fitting)を交互に行う算法である。この手法は、複数のサブユニットを用いて、マップ全体を覆うような配置を効率的に探索できる。しかし、実際の問題では、マップの一部にか原子モデルを用意できない場合もある。こうした場合に対応するように、サブユニットの周辺にマスク領域を設け、その領域内のみでセグメンテーションを行う「マスク付きセグメンテーション・フィッティング法」（マスク付きセグ・フィット法）も開発した。この方法は、ローカルなフィッティング問題に対応できる。

さらに、全体の計算手続きを I：初期化、S:探索、R:精密化の 3 ステップでサブユニットの配置探索を行うようプログラムを整理した。前述の「マスク付きセグ・フィット法」は S-ステップの一つに対応する。R-ステップは、最急降下法を採用している。初期配置生成の I-ステップでは、これまでランダム法と主成分軸法を実装していたが、より包括的に探索を行うために、「格子配置法(Grid-Layout)」による初期配置の生成法の実装を行った。この方法では、マップ、サブユニットを粗い格子空間に並進配置、回転角度を離散的に試すことで、包括的に探索を行う。しかし、この方法は原則として１サブユニットにしか適応できない。より多数のサブユニット群をマップに配置する探索を行う場合、各サブユニットの探索を別途行い、その探索結果を組み合

わせて複数サブユニットの初期配置を作る「単サブユニット検索組み合わせ方法（CSS; Combination of Single-subunit Searches)」も開発し、実装した。これらの方法により、これまでより効率的に複数サブユニットの探索を行うことが可能となった。



以下に改良されたプログラム *gmfit* を用いて、GroEL/ES の電顕マップ(EMD-1024,分解能 23.5Å)に、21 個のサブユニットの原子モデルを重ねる計算を行った試験計算例を示す。別途決定された X 線結晶構造に 7.88Åの構造が 1CPU で 86 分の計算時間で得られた。同様の計算を 2008 年の *gmfit* で行った場合は、1CPU で 160 時間を費やして 14.7Åの RMSD であったと報告されており、計算効率が顕著に改善されたことがわかる。



C7の拘束付き　I-step:CSS法(３種の各サブユニットについて、格子配置法、マスク付きセグフィット法、最急降下法)
S-step:セグフィット法　R-step:最急降下法

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Takeshi Kawabata	４．巻 16
２．論文標題 Detection of cave pockets in large molecules: Spaces into which internal probes can enter, but external probes from outside cannot	５．発行年 2019年
３．雑誌名 Biophysics and Physicobiology	６．最初と最後の頁 391-406
掲載論文のDOI（デジタルオブジェクト識別子） 10.2142/biophysico.16.0_391	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 川端 猛.	４．巻 59
２．論文標題 混合正規分布を用いた剛体フィッティング法：電顕密度マップに原子モデルを重ねる計算	５．発行年 2019年
３．雑誌名 生物物理	６．最初と最後の頁 320-323
掲載論文のDOI（デジタルオブジェクト識別子） 10.2142/biophys.59.320	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kawabata Takeshi	４．巻 203
２．論文標題 Gaussian-input Gaussian mixture model for representing density maps and atomic models	５．発行年 2018年
３．雑誌名 Journal of Structural Biology	６．最初と最後の頁 1-16
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jsb.2018.03.002	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 川端　猛
２．発表標題 大きな分子の「空洞ポケット」：小さな球は入れるが、大きな球は外からは入れない空間
３．学会等名 第19回日本蛋白質科学会年会 第71回日本細胞生物学会大会合同年次大会
４．発表年 2019年

１．発表者名
川端　猛、中村春木、栗栖源嗣

２．発表標題
マスク付きセグメンテーション・フィット法による複数サブユニットの電顕マップへの局所重ね合わせ

３．学会等名
第57回日本生物物理学会年会

４．発表年
2019年

１．発表者名
川端　猛、栗栖源嗣

２．発表標題
EM informatics: archiving raw 2D images and fitting atomic models into a map

３．学会等名
第57回日本生物物理学会年会

４．発表年
2019年

１．発表者名
川端　猛、中村春木、栗栖源嗣

２．発表標題
サブユニット原子モデル群を電顕密度マップの一部に重ね合わせる算法の開発

３．学会等名
第18回日本蛋白質科学会年会

４．発表年
2018年

１．発表者名
川端　猛、中村春木、栗栖源嗣

２．発表標題
マスク付きガウス関数による電顕３次元密度マップ内のαヘリックスを認識する手法の開発

３．学会等名
第56回日本生物物理学会年会

４．発表年
2018年

１．発表者名
Kawabata T, Suzuki H, Kurisu G

２．発表標題
Databases and Web services from PDBj for Electron Microscopy

３．学会等名
Asian Crystallographic Association Conference (AsCA 2018)（国際学会）

４．発表年
2018年

１．発表者名
Kawabata T, Suzuki H, Nakamura H

２．発表標題
gmfit : an approximated 3D shape for atomic model and density map using Gaussian Mixture model

３．学会等名
IUPAB & EBSA 2017（国際学会）

４．発表年
2017年

１．発表者名
川端　猛、中村春木

２．発表標題
ガウス関数入力型混合正規分布モデルによる３次元密度マップの近似表現

３．学会等名
第55回日本生物物理学会年会

４．発表年
2017年

１．発表者名
Kawabata T, Nakamura H.

２．発表標題
Helix detection and subunit fitting using Gaussian mixture model.

３．学会等名
CryoEM Structure Challenges Workshop（国際学会）

４．発表年
2017年

〔図書〕　計1件

1．著者名 Takeshi Kawabata	4．発行年 2018年
2．出版社 Springer	5．総ページ数 17
3．書名 Chapter 14 "Rigid-Body Fitting of Atomic Models on 3D Density Maps of Electron Microscopy" in the book "Integrative Structural Biology with Hybrid Methods"	

〔産業財産権〕

〔その他〕

フィッティングプログラムgmfitのページ
https://pdbj.org/gmfit/
ポケット同定プログラムghecomのページ
https://pdbj.org/ghecom/

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