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　 ８ 日現在

機関番号：８２５０２
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ０６１０１

研究課題名（和文）立体構造データベースと電子顕微鏡画像を用いた立体構造モデル構築手法の開発

研究課題名（英文）Development of 3D model construction method using structure database and electron microscope images

研究代表者
松本　淳（Matsumoto, Atsushi）

国立研究開発法人量子科学技術研究開発機構・量子生命科学研究所・主幹研究員

研究者番号：１０３９９４２０

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究では、生体分子の電子顕微鏡画像が与えられた際に、それに合致あるいは類似する立体構造を、立体構造データベースに登録された生体分子の中から選び出す計算機手法を開発した。その際、代表者が開発した疑似電顕画像作成手法とニューラルネットワークを組み合わせて用いた。
現在、データベースに登録されている約２０万個の生体分子の構造のうち、電子顕微鏡で観測されるような大きな構造（約２万個）から電顕画像を作成し、機械学習用のデータセットとした。立体構造の大きさによってクラス分けを行い、各クラスで機械学習を行った。最も大きな構造クラスでは、正答率が約７割（上位３位までに正解が含まれるのは約９割）だった。

研究成果の学術的意義や社会的意義
本研究で開発した計算機手法を用いることにより、立体構造が分かっていない生体分子であっても、それに類似する構造がデータベースに登録されていれば、その電子顕微鏡画像をもとに、立体構造を速やかに類推することができる。
電子顕微鏡による立体構造解析では、通常クライオ電子顕微鏡が用いられるが、非常に高価で、これを持つ施設も限られている。本手法を用いることで、従来の電子顕微鏡を活用して、立体構造解析を行うことができる。

研究成果の概要（英文）：In this study, a computer method to select or find a matching or similar three-dimensional structure from a database of biomolecules when given an electron microscope image of a biomolecule was developed. The method used a combination of the microscopy image creation technique developed by the lead author and neural networks. For this purpose, a dataset of electron microscopy images was created for machine learning from approximately 20,000 large biomolecules selected from the current database consisting of approximately 200,000 biomolecules. The structures were classified by size and machine learning was performed in each class. In the largest structure class, the accuracy rate was about 70% (with the correct answer included in the top three answers about 90% of the time).

研究分野：生物物理

キーワード：プログラム開発　ニューラルネットワーク　電子顕微鏡　生体分子　立体構造

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年、クライオ電子顕微鏡（以下クライオ電顕）と 3 次元再構成技術（多数の 2 次元電子顕微鏡像から 3 次元立体構造を構築する計算技術）により、生体分子の 3 次元立体構造（3D-EM 構造）が多数解かれている。そして、3D-EM 構造に対し、計算機手法を用いることにより、原子分解能のモデル構造の構築が行われる。

クライオ電顕による構造解析では、X線結晶構造解析法と違い、サンプルを結晶化する必要がないという大きな利点があるが、いくつかの問題がある。まず、クライオ電顕では、従来のネガティブステイン（負染色剤）による電子顕微鏡観察よりも大量のサンプルを必要とするため、サンプルの準備に手間と費用がかかる。また、原子分解能を達成できるようなクライオ電顕は極めて高価なうえに、多額の運用費が必要である。さらに、3D-EM 構造の構築には、膨大な枚数の電顕画像を処理するための高性能の計算機システムが必要であり、導入には、電子顕微鏡ほどではないにしても、多額の費用が必要である。以上のように、3D-EM 構造の構築には、手間と費用がかかり、そのハードルはかなり高い。実際、日本国内で、これを行える研究施設は限られている。

代表者は、これまでに、生体分子のネガティブステイン電顕画像と結晶構造をもとに、原子モデル構造の構築を行う計算機手法（2D ハイブリッド解析法）を開発した（Matsumoto, et al., Sci. Rep., 2017）。この手法では、結晶構造（あるいはモデル構造）を初期構造にして、計算機シミュレーション手法により様々な変形構造を作成し、さらに、それぞれの変形構造をもとに多くのシミュレート電顕画像を作成して、実際の電顕画像に合致する変形構造の探索を行う。

この手法を開発した第１の理由は、構造変化しやすい生体分子の 3D-EM 構造構築は容易ではないため、3D-EM 構造構築が不要の原子モデル構築手法が必要だったからであるが、結果的には、従来の電子顕微鏡画像に対しても使える計算機手法となり、従来の電子顕微鏡を使っている多くの研究者の関心を引くこととなった。

２．研究の目的

本研究の目的は、生体分子の電子顕微鏡画像が与えられた際に、それに合致あるいは類似する立体構造を、立体構造データベースに登録された生体分子の中から選び出す計算機手法あるいは AI（人工知能）を開発することである。これにより、立体構造が分かっていない生体分子であっても、それに類似する構造がデータベースに登録されていれば、その電子顕微鏡画像をもとに、立体構造を速やかに類推することができる。

３．研究の方法

本研究では、2 種類の計算を行う。一つは、PDB(Protein DataBank:タンパク質や核酸など生体高分子の 3 次元構造の原子座標を蓄積しているデータベース)に登録されている構造をもとにして、電顕画像を作成する計算で、もう一つは、作成された膨大な数の電顕画像を用いて、計算機に学習させる計算（機械学習）である（図１参照）。

一つ目の計算では、代表者が開発した“2D ハイブリッド解析法”のうち、PDB 構造から電顕画像を作成する部分を用いた。ただし、本研究では、膨大な数の電顕画像を機械学習に用いるため、計算機プログラムを改良し、高速に画像を作成できるようにした。

二つ目の計算は、与えられた電顕画像がどの構造のものであるかを識別する AI を構築するためのものであり、できるだけ高い正答率（accuracy）を得るために、様々な構造のニューラルネットワークを構築した。

４．研究成果

現在、PDB には、約２０万個の生体分子の立体構造が登録されているが、そのうち、電子顕微鏡で観測されるような、ある程度大きな構造（約２万個）に対象を絞った。そして、立体構造情報ごとに、電子線の照射方向（投影方向）やネガティブステインの厚み、それにピクセルサイズ（画像の１画素の幅が、何オングストロームに対応するか）の異なる数千枚の疑似電顕画像を作成した。これにより、機械学習で用いるための、合計約１億枚の電顕画像からなるラベル付きデータセットを構築した。ここでのラベルは、生体分子の各構造に付与された PDB の ID である。

ただ、PDB には形状が類似した構造が多く登録されていて、それらの電顕画像を区別することは難しいので、機械学習において正答率が上がらないことにつながる。そのため、登録された構造を、形状の類似度でグループ分けし、各グループの代表構造の電顕画像をデータセットとして用いることにした。このグループ分けの作業のために、各構造の質量・慣性モーメントなどの物理量を計算するとともに、２つの構造を重ね合わせることで構造の類似度を計算する計算機手法の開発・適用を行った。

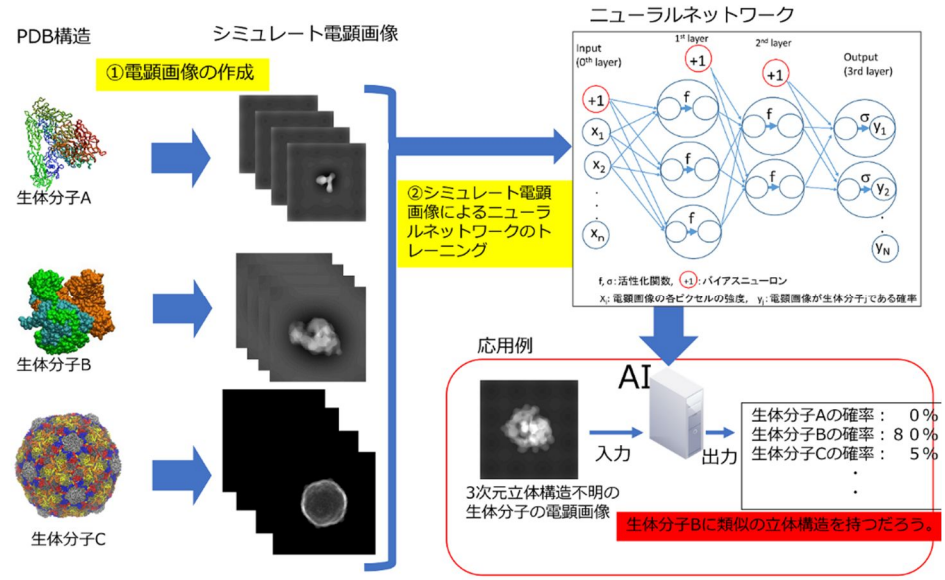


**図１：本研究の概要。生体分子の電子顕微鏡画像をもとに、その立体構造を類推する計算機手法（応用例の部分）を開発する。立体構造データベースに登録された生体分子に２D ハイブリッド解析法を適用してシミュレート電顕画像を作成し、それをニューラルネットワークのトレーニングに用いることで、生体分子の電顕画像を識別する AI（人工知能）を構築する。**

ニューラルネットワークの入力には、一定の大きさの画像を用いる（例えば、224ｘ224 ピクセル）。そして、分子の大きさは、分子を識別する重要な情報なので、学習の際には、すべての画像のピクセルサイズは一定にすべきである。そのため、一度にすべての電顕画像を用いて学習を行おうとすると、大きな分子が画像に収まるように大きなピクセルサイズを用いる必要があるので、小さな分子は画像上で小さくなり、その情報量は少なくなる。また、大きさが明らかに違う分子の電顕画像を区別することは容易なので、それらを同時に学習させる必要性は少ない。そこで、PDB 構造を大きさでクラス分けして、各クラスの電顕画像に対して、別々に機械学習を行った。

正答率の高い AI を得るために、様々なニューラルネットワークを構築した。研究開始当初は、CNN（畳み込みニューラルネットワーク）をゼロから構築し、訓練をしていたが、正答率はあまり高くならなかった。そこで、学習済みネットワークを利用すること（転移学習）にした。VGG16, VGG19, InceptionV3, Resnet50 などの様々な学習済みネットワークを試した。さらに、正答率を上げるためと過学習（学習データに対しては高い正答率を出すが、未知データに対しては同様の正答率を出せないモデルが構築されてしまうこと）を避けるために、dropout 層や L1/L2 正則化層などをネットワークに追加し、それらの効果を調べた。その結果、学習済みネットワークとして Resnet50 を利用し、さらに、dropout 層と L2 正則化層を追加した場合に、高い正答率が得られ、過学習を避けることができた。

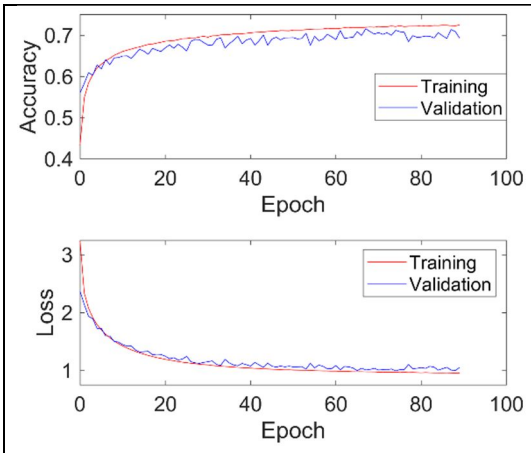


**図２：立体構造の大きさが最大のクラスの電顕画像の学習結果。上が正答率（accuracy）、下が損失（loss）の推移を示す。損失が小さくなるように、学習は行われる。**

図２には、立体構造の大きさが最も大きな 100 個の構造グループから成るクラスの電顕画像の学習結果を示している。作成した電顕画像のデータセットは、学習(Training)用、検証（Validation）用、テスト（Test）用に分け、学習の際には学習用データセットのみを用いた。そして、各エポック(Epoch)終了時に、検証データセットを用いて、モデルの検証を行った。ここでエポックとは、学習の繰り返し回数に関係する数であり、１エポックで、学習用画像の枚数分の回数、繰り返し学習を行ったことになる。さらに、損失が最小となった時点で、テスト用データセットを用いてテスト計算を行ったところ、正答率（最も可能性が高いと AI が判断した構造が正解の確率）は 70.6%だった。さらに、上位３位までに正解が含まれていた場合も含めると、正答率は 89.6%だった。

本成果は、第６０回日本生物物理学会年会で発表した。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　1件／うちオープンアクセス　2件）

| | |
|---|---|
| １．著者名<br>Matsumoto Atsushi、Sugiyama Masaaki、Li Zhenhai、Martel Anne、Porcar Lionel、Inoue Rintaro、Kato Daiki、Osakabe Akihisa、Kurumizaka Hitoshi、Kono Hidetoshi | ４．巻<br>118 |
| ２．論文標題<br>Structural Studies of Overlapping Dinucleosomes in Solution | ５．発行年<br>2020年 |
| ３．雑誌名<br>Biophysical Journal | ６．最初と最後の頁<br>2209～2219 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1016/j.bpj.2019.12.010 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>該当する |

| | |
|---|---|
| １．著者名<br>Matsumoto Atsushi | ４．巻<br>16 |
| ２．論文標題<br>Dynamic analysis of ribosome by a movie made from many three-dimensional electron-microscopy density maps | ５．発行年<br>2019年 |
| ３．雑誌名<br>Biophysics and Physicobiology | ６．最初と最後の頁<br>108～113 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.2142/biophysico.16.0_108 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

| |
|---|
| １．発表者名<br>松本淳 |
| ２．発表標題<br>Deep learning of computer-generated electron microscopy images to identify biomolecules |
| ３．学会等名<br>第６０回日本生物物理学会年会 |
| ４．発表年<br>2022年 |

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（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 |
| --- | --- |