生物物理 62（3），193-197（2022）
DOI: 10.2142/biophys.62.193
受理日：2022 年 3 月 1 日

理論／実験 技術

深層学習技術を用いたクライオ電子顕微鏡データに潜むタンパク質運動性情報の抽出

松本篤幸[1]，寺山　慧[2]，奥野恭史[1]
[1] 京都大学大学院医学研究科
[2] 横浜市立大学生命医科学研究科

1. はじめに

生命現象はタンパク質などの生体高分子の働きの上に成り立っている．そのため生命現象の分子メカニズムを知る上でタンパク質の機能を正しく理解することは重要である．タンパク質の機能はその立体構造とその運動性によって精密に制御されている．このことから，タンパク質分子の高分解能立体構造解析すなわち構造生物学が生命科学の中核の 1 つとなっている．

高分解能の立体構造を決定するための主な実験的手法として，核磁気共鳴（NMR）法，X 線結晶構造解析，低温電子顕微鏡（cryo-EM）単粒子解析が挙げられる．このうち特に cryo-EM は，近年のハードウェアと解析技術の発展[1]により，100 kDa 以下の比較的小さな生体高分子のみならず巨大かつ複雑な生体高分子の立体構造を次々と解明することで今日の分子生物学の発展に多大な貢献を果たしている．一方，分子の運動性解析は NMR 法や水素–重水素交換質量分析法（HDX-MS）などの実験的手法に加え分子動力学（MD）計算によって取り組まれてきた．これらの手法は生体高分子の動的振る舞いを高分解能で定量的に計測可能であるが，それらの手法を巨大かつ複雑な生体高分子に適用するためには原理的に多くの困難が伴う．

cryo-EM 単粒子解析では，透過型電子顕微鏡による試料撮影で得られる様々な方位からの大量の生体高分子画像（単粒子画像）を収集し，再構成することで 3 次元密度マップを得る（**図 1**）．撮影試料はタンパク質溶液を瞬間的に凍結して準備するため，得られる単粒子画像中には溶液中で見られる様々な構造状態が含まれる．すなわち，それらの単粒子画像によって再構成された 3 次元密度マップの中には溶液中での動的振る舞いに関わる情報が潜んでいるといえる．これを反映して，タンパク質分子内部で疎水性コアを形成しているような「固い」領域の密度マップは構造の均一性によりはっきりと見え（＝強度が強い），分子表面に露出しているループ領域などの「柔軟な」領域の密度マップは様々な構造状態が平均化されてしまうためぼやけている（＝強度が弱い）（**図 1**）．このように密度マップ強度と運動性の間に関連性があるということは経験的に知られている一方，密度マップ強度は運動性以外にも試料調製過程での局所的な変性やグリッド上での分子の向きの偏りなどの複数の要因による影響を受けるため，単純な密度マップ強度からの運動性推定は不可能であった．

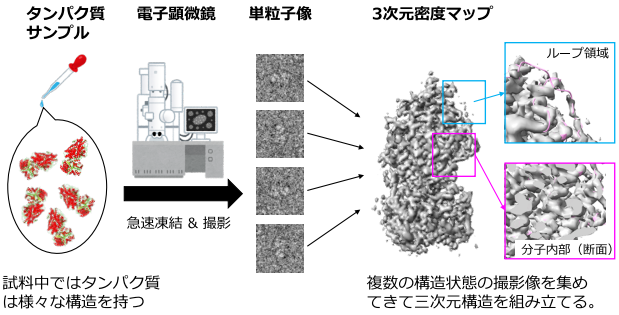


図 1
運動性に依存した構造多様性に由来する 3 次元密度マップ強度の違いの様子．「固い」分子内部領域（マゼンタの四角で拡大図を示す）でははっきりとした密度マップが得られている．一方，「柔軟な」ループ領域（シアンの四角で拡大図を示す）ではところどころ密度マップが途切れている．

本稿では，近年発展著しい深層学習技術を用いて，cryo-EM の 3 次元密度マップ強度情報から直接的にそこに潜む運動性情報を抽出する手法 Dynamics Extraction From cryo-em Map（DEFMap）[2]について紹介すると共にその展望と現時点での課題について議論する．

Extraction of Protein Dynamics Hidden in Cryo-EM Maps Using Deep Learning
Shigeyuki MATSUMOTO[1], Kei TERAYAMA[2] and Yasushi OKUNO[1]
[1]Graduate School of Medicine, Kyoto University
[2]Graduate School of Medical Life Science, Yokohama City University

2. 3 次元畳み込みニューラルネットワーク（3D-CNN）

DEFMap では深層学習技術の 1 つである 3 次元畳み込みニューラルネットワーク（3D-CNN）を利用することで 3 次元密度マップからの運動性情報の抽出を実現している．畳み込みニューラルネットワークモデルは予測に重要な入力データの局所的な特徴・パターンを捉えることが可能で，画像認識や音声認識において高い性能を示すことが知られている．これを 3 次元に拡張した 3D-CNN モデルでは空間的特徴の学習が可能であり，コンピュータ断層撮影や核磁気共鳴イメージングなどにおける 3 次元オブジェクトの検出やクラス分類に広く利用されている [3]．

3. 学習データセットの生成と学習

DEFMap では cryo-EM の 3 次元密度マップの強度パターンとその運動性情報との関係性を学習している．これを実現するためには，密度マップと運動性情報が紐づいた大規模なデータセット（学習データ）が必要である．日々決定されている cryo-EM の 3 次元密度マップ及びそこから得られる原子モデルはデータベース Electron Microscopy Data Bank（EMDB）並びに Protein Data Bank（PDB）に蓄積されている．一方，それらの生体高分子の運動性に関するデータベースは存在せず，また実験的に計測することは非現実的なため，MD 計算を用いて運動性情報を取得することとした．

学習対象のタンパク質として，①比較的簡便に MD 計算を実施することのできるタンパク質であること②4.5 Å よりも良い分解能で 3 次元密度マップが決定されていることの 2 点を指標に選抜を行い，結果的に 25 種類のタンパク質を学習に用いた．

3 次元密度マップについて，効率的な学習を行うための前処理としてマップの全体分解能が 5 Å になるように low-pass filter を適用した．DEFMap では局所の運動性予測を実現するというアイデアの下，low-pass filter 適用後の各 3 次元密度マップから重原子位置を中心に局所の密度マップデータを 1 辺 15 Å の立方体で切り出し，入力データ（sub-voxel）とした（図 2a 上段）．3 次元密度マップ中の重原子位置は，密度マップから構築された原子モデルの（x, y, z）座標から最も近い位置のグリッドとした．sub-voxel を学習データにすることにより大量の局所密度マップ環境を学習データとして準備することが可能であり，これらのデータを用いて学習したモデルについて多様な外部データへの汎用性が期待できる．最終的に本研究では 4,249,300 個の sub-voxel を学習データとして使用した．一方，運動性情報

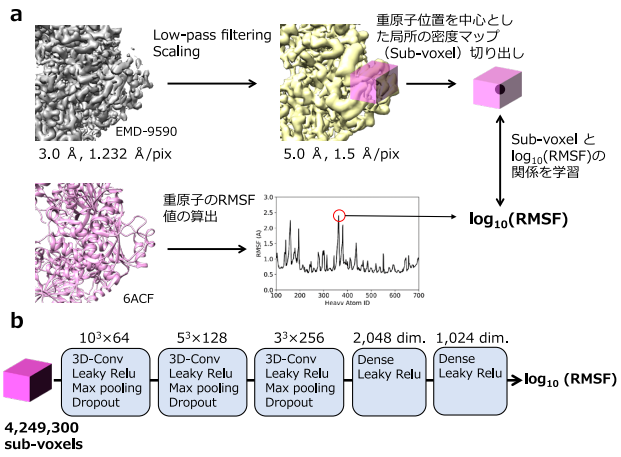


図 2
DEFMap における学習．（a）学習データセットの準備．（b）DEFMap における深層学習モデル構成の概要．

としては 30 nsec の MD 計算結果から算出した重原子の root-mean square fluctuation の常用対数値 $\log_{10}$（RMSF）を用いた（図 2a 下段）．MD 計算の入力初期座標はそれぞれの 3 次元密度マップから構築された原子モデルから準備し，MD エンジンとして GROMACS 2016.5[4]を利用した．以上の手順で準備した各 sub-voxel（説明変数）と運動性情報（目的変数）との関係性を図 2b に示すニューラルネットワークモデルで学習した．

4. 運動性予測の性能検証

25 種類の学習タンパク質のうち 1 つをテストデータとし，残り 24 種を教師データとした交差検証（Leave-one-out cross validation 法）により DEFMap の運動性予測性能を評価した．その結果，運動性との相関係数 r の平均（±分散）は 0.665（±0.124）であった．一方，密度マップ強度そのものと運動性との相関係数 r は 0.459（±0.179）であった．このことは，深層学習モデルを用いることで密度マップ強度から目的とする空間的パターンの抽出に成功していることを示している．図 3a には交差検証の評価のうちの一例を示している．

得られたモデルの外部データ（学習に利用していない密度マップ）に対する性能を評価するため，EMDB 及び PDB より新たに EMD-4241/6FE8[5]，EMD-7113/6BLY[6]，EMD-20308/6PCV[7]の 3 種類の cryo-EM データを取得し，DEFMap による予測結果と MD 計算で得た運動性を比較したところ，相関係数 r がそれぞれ 0.727，0.748，0.711 と良い一致が見られた．図 3b にはそのうち EMD-20308/6PCV における比較結果，並びに予測結果を立体構造上へマッピングした様子を示しており，DEFMap が分子内部と溶媒露出表面の運動性

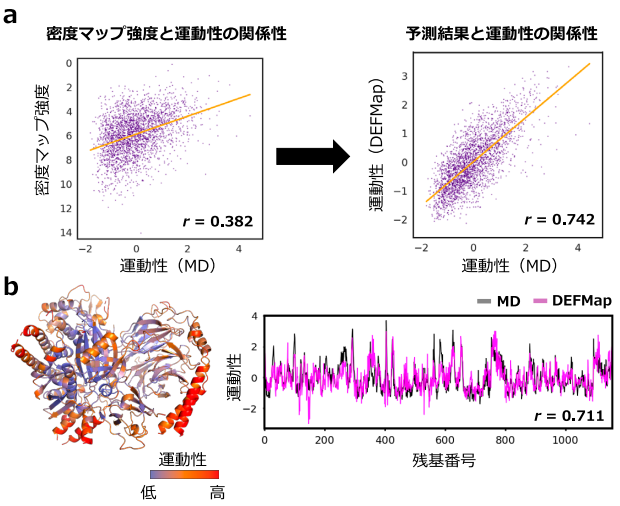


図 3
外部データを用いた DEFMap の予測性能評価．（a）MD 計算で得た運動性と密度マップ強度（左）並びに DEFMap での予測結果（右）との関係．それぞれの値は残基ごとの平均値を標準化して用いている．（b）外部データに対する運動性予測結果の立体構造上へのマッピング（左）と MD 計算で決定した運動性との比較（右）．文献 2 の図を改変．

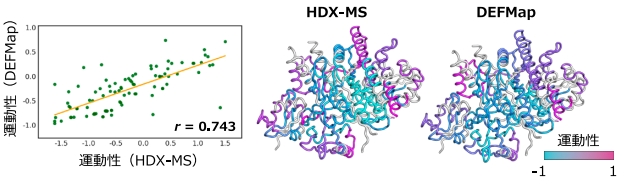


図 4
実験データを用いた DEFMap の予測性能評価．予測結果は HDX-MS で検出されたペプチドフラグメントの結果に従ってフラグメントごとの平均を計算し，標準化して比較を行っている．左にはそれぞれのペプチドフラグメントの相関を，右にはそれらを立体構造上にマッピングした様子を示している（高い運動性と予想された領域を紫色傾向で示す）．文献 2 の図を改変．

の違いなど立体構造上の一般的な特徴を良く捉えている様子が観察できる．

DEFMap は MD 計算で得られる運動性と密度マップ強度との関係性を学習したモデルである．このことから，予測結果を実験的に測定された運動性と比較することは重要である．EMD-20308/6PCV について，予測結果を HDX-MS 法で決定された運動性と比較したところ，両者の間では良い相関が見られていた（$r = 0.743$，**図 4**）．この比較結果は DEFMap の予測結果により運動性の議論が可能なことを支持するものである．

5. 構造生物学的研究への貢献

では密度マップから直接的に運動性を予測できたとして，構造生物学的研究においてどのような貢献を期待できるのであろうか．本章では筆者らが実際に

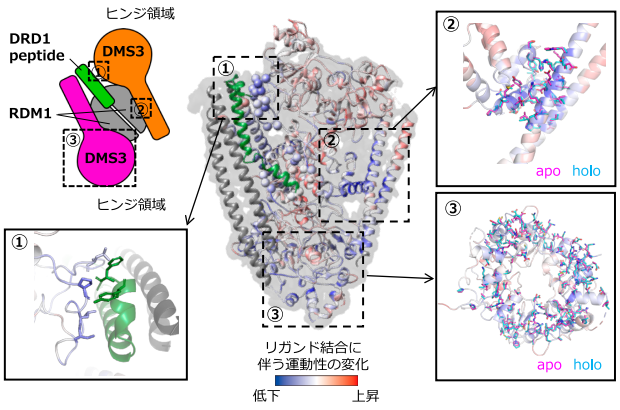


図 5
リガンド結合に伴う運動性変化の可視化．原子モデルは holo 状態の予測結果から apo 状態の予測結果を引いた値で色分けされており，青傾向の領域がリガンド結合に伴って運動性が抑制されたことを示している．またリガンドは緑色で示している．リガンド結合部位の拡大図（①）ではリガンド認識に重要な残基をスティック表示で示している．また RDM1-DMS1 相互作用界面（②）並びに DMS1 ヒンジ領域の拡大図（③）には運動性の抑制が予測された残基をスティック表示で示している．文献 2 の図を改変．

DEFMap で見出した分子メカニズム解析研究の指針となり得る興味深い知見について紹介する．

一般的にリガンドの相互作用に伴ってその結合部位は安定化され，運動性が低下する．DNA メチル化導入と関連する meristem silencing 3（DMS3）–RNA-directed DNA methylation 1（RDM1）複合体を対象に，defective RNA-directed DNA methylation 1（DRD1）ペプチドが結合した 3 次元密度マップ（EMD-20081，holo 状態）と非結合型の 3 次元密度マップ（EMD-20080，apo 状態）[8]それぞれについて予測した運動性を比較してみると，リガンド結合に伴い，DRD1 ペプチド認識に重要な残基を中心とした結合部位の運動性の低下を検出することができた（**図 5-①**）．興味深いことに DRD1 ペプチド結合に伴う運動性抑制は，その結合部位から遠くに位置している RDM1-DMS3 相互作用界面（**図 5-②**）並びに DMS3 のヒンジ領域（**図 5-③**）においても観察された．このことはペプチド結合の影響が分子内を伝わり，複合体形成並びに DMS3 の安定化を誘導していることを示唆している．特筆すべき点として，密度マップに基づいて構築された apo 状態及び holo 状態の原子モデルではこれらの領域における明確な差が認められなかった（**図 5-②, ③**）．このことは，これらの運動性の変調は通常の構造解析過程では見過ごされてしまう恐れがあることを示しており，DEFMap の構造生物学的研究における有用性を強調するものである．

DEFMap による運動性解析の優位性として，1．入力が密度マップであるため分子量による制限を受けない，

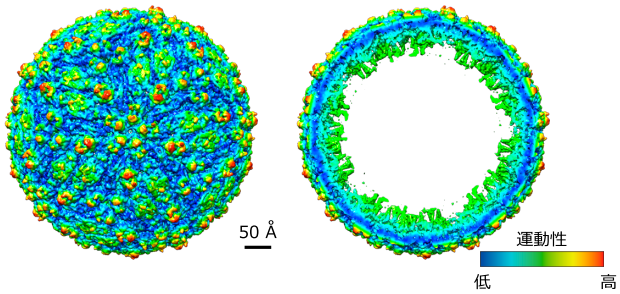


図 6
ジカ熱ウイルス粒子の 3 次元密度マップ（EMD-8139）を用いた運動性予測．文献 2 の図を改変．

2．原子モデルを必要としないという 2 点が挙げられる．例えば通常の実験的計測技術で巨大なウイルス粒子の運動性を解析する場合，シグナルの重なりなどその巨大さに由来する様々な障壁が存在する．また MD 計算の実施には高い計算コストを必要とする上，高分解能の原子モデルが得られていない場合には信頼性のある結果を得ることは難しい．DEFMap での予測においてはこれらの困難とは無縁であるため，3 次元密度マップが得られていれば簡便に運動性解析を行うことができる．**図 6** には実際に cryo-EM 単粒子解析によって得られた巨大ウイルス粒子の 3 次元密度マップ[9]に対する運動性予測を実施した例を示している．

6. DEFMap の現モデルの限界と高度化

密度マップの全体分解能に対する予測性能の依存性を検証したところ，分解能 7 Å 付近を境に低分解能側で予測性能が低下することを確認している．これは低分解能密度マップには予測に必要なだけの情報量が不足していることを示唆している．Cryo-EM で得られる密度マップでは局所分解能が分子内で幅広く異なることから，局所分解能が極端に悪い領域については DEFMap の予測結果を慎重に吟味する必要がある．

また機械学習技術における別の限界として，学習データに入っていない入力データに対する予測が難しい点が挙げられる．DEFMap の現時点でのモデルでは MD 計算で運動性データを生成する都合上，比較的分子量の小さい可溶性タンパク質を学習データとして用いた．そのため膜タンパク質の膜貫通領域などの特殊な環境下の密度マップデータは学習データセットに入っていない．このことから，現時点で界面活性剤などの特殊な密度マップを含む sub-voxel に対する高い予測精度は期待できない．

前者の限界については cold field emission guns[9]などの測定技術並びに画像解析技術の向上に伴って日々分解能が改善されている現状を鑑みると，近い将来解消されることが十分に期待できる．一方後者については今後の学習データの拡充による解決が期待できる．DEFMap における学習データ拡大のボトルネックは大規模な MD 計算である．莫大な scalability を持つ「富岳」の公開を始めとしたスーパーコンピューター利用環境の整備がこの点の解決を強力に後押ししてくれると考えられる．

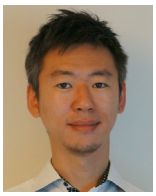
7. まとめ

DEFMap は cryo-EM 単粒子解析法によって得られる 3 次元密度マップ全体を直接解釈することで分子全体の運動性の可視化を可能にする．これにより通常の実験的アプローチでは解析に困難が伴うような巨大かつ複雑な生体高分子の運動性の解析を簡便に実施できる．本稿で紹介したようなリガンド結合に伴う遠位の運動性変化の検出などを通じて，DEFMap の利用が cryo-EM 単粒子解析に基づく分子メカニズム解明を加速することを期待している．また DEFMap では MD 計算をビッグデータ生成に利用することで実験データと深層学習技術の融合的な研究を実現した．本研究がこの新たな融合的研究アプローチの発展において先駆け的研究になると考えている．

最後に，DEFMap のコードは github 上で公開しており（https://github.com/clinfo/DEFMap），深層学習用 Python ライブラリ TensorFlow, Keras, 分子解析用 Python ライブラリ HTMD, cryo-EM 解析用プログラム EMAN2（いずれも無償で学術利用可能）が使える環境を準備することで誰でも利用可能である．

文　献

1) Cheng, Y. (2018) Science **361**, 876-880. DOI: 10.1126/science.aat4346.
2) Matsumoto, S. *et al.* (2021) Nat. Mach. Intell. **3**, 153-160. DOI: 10.1038/s42256-020-00290-y.
3) Ji, S. *et al.* (2013) IEEE Trans. Pattern Anal. Mach. Intell. **35**, 221-231. DOI: 10.1109/TPAMI.2012.59.
4) Abraham, M. J. *et al.* (2015) SoftwareX **1-2**, 19-25. DOI: 10.1016/j.softx.2015.06.001.
5) Zhang, W. *et al.* (2018) Cell Rep. **24**, 744-754. DOI: 10.1016/j.celrep.2018.06.068.
6) Sun, Y. *et al.* (2018) Proc. Natl. Acad. Sci. U.S.A. **115**, E1419-E1428. DOI: 10.1073/pnas.1718723115.
7) Cash, J. N. *et al.* (2019) Sci. Adv. **5**, eaax8855. DOI: 10.1126/sciadv.aax8855.
8) Wongpalee, S. P. *et al.* (2019) Nat. Commun. **10**, 3916. DOI: 10.1038/s41467-019-11759-9.
9) Kato, T. *et al.* (2019) Microsc. Microanal. **25**, 998-999. DOI: 10.1017/S1431927619005725.


松本篤幸

松本篤幸（まつもと　しげゆき）
京都大学大学院医学研究科特定准教授
2009 年大阪大学大学院薬学研究科博士課程修了，博士（薬学）．
研究内容：構造生物学
連絡先：〒 606-8507 京都市左京区聖護院河原町 54
E-mail: matsumoto.shigeyuki.4z@kyoto-u.ac.jp


寺山　慧

寺山　慧（てらやま　けい）
横浜市立大学大学院生命医科学研究科准教授
2016 年京都大学大学院人間･環境学研究科博士課程修了，博士（人間・環境学）．
研究内容：機械学習の生命科学・材料科学・水産海洋工学への応用
連絡先：〒 230-0045 神奈川県横浜市鶴見区末広町 1-7-29 A414
E-mail: terayama@yokohama-cu.ac.jp


奥野恭史

奥野恭史（おくの　やすし）
京都大学大学院医学研究科教授
2001 年京都大学大学院薬学研究科博士課程修了，博士（薬学）．
研究内容：創薬計算科学，ビッグデータ医科学
連絡先：〒 606-8507 京都市左京区聖護院河原町 54
E-mail: okuno.yasushi.4c@kyoto-u.ac.jp

理論／
実験 技術