連載

5分で分かる!? 有名論文ナナメ読み

富井健太郎（産業技術総合研究所）

John Jumper et al. :
Highly Accurate Protein Structure Prediction with AlphaFold

Nature 596(7873), 583-589 (2021)

タンパク質のカタチ

「タンパク質」にどのようなイメージをお持ちだろうか．三大栄養素の1つであることからも容易に推量できるように，タンパク質は生物を構成する主要な要素の1つであり，ヒトの場合，皮膚や毛髪，筋肉のみならず，昨今耳にする機会も多い抗体に至るまで，タンパク質で構成されている．

タンパク質は，さまざまな物理化学的性質を持つ（主に）20種類のアミノ酸と呼ばれる構成単位が鎖状に連なった高分子であり，タンパク質の種類ごとに固有のアミノ酸「配列」を有している．例外もあるが，多くのタンパク質は，生理的条件下で天然状態と呼ばれる固有の立体構造をとって機能している．天然状態は，タンパク質の種類ごとに実にさまざまな構造である（図-1）が，1972年ノーベル化学賞を受賞したAnfinsenの研究☆1により，天然状態を規定する情報がタンパク質のアミノ酸配列に書き込まれていることが示された．以来，アミノ酸配列からのタンパク質立体構造予測は，生物学上の重要な課題の1つである．

☆1 https://www.nobelprize.org/prizes/chemistry/1972/summary/

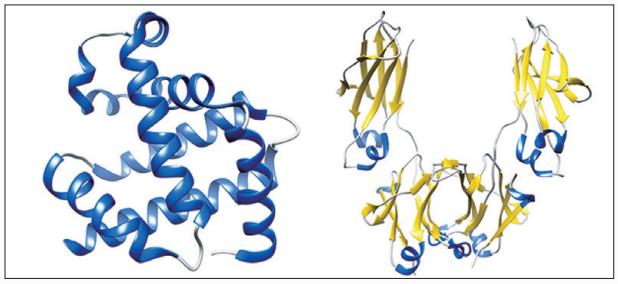

図-1 ヒトの持つタンパク質の例．左）ミオグロビン［PDB* ID：3rgk］，右）免疫グロブリン（抗体）［PDB* ID：4wi6］（* https://www.wwpdb.org）

長年にわたるタンパク質立体構造観測データの蓄積や計算生物学の発展に伴い，タンパク質立体構造予測法の性能を客観的に評価するための国際的な実験CASP（Critical Assessment of Techniques for Protein Structure Prediction）が1994年から隔年開催されている．昨年開催された第14回目のCASP14☆2において，attentionを利用した画期的な手法AlphaFold 2（紹介論文の表記に従い，以降一部を除きAlphaFoldと表記）が発表され，今年になりその詳細が明らかになった．この手法では，立体構造を予測したいタンパク質のアミノ酸配列と，後述するMSAおよび鋳型構造から計算される特徴量を入力とし，立体構造（水素を除く原子座標）とアミノ酸単位での予測精度が計算される．以下ではまず，この手法についての理解を深めるため，いくつかの予備知識を紹介する．

MSA

現在自然界に存在するタンパク質は，長い進化の過程を経て祖先タンパク質から派生してきたものと考えられる．たとえば，図-1に示したミオグロビンという種類のタンパク質1つをとっても，同様の働きをする似たタンパク質がほかの生物種にも広く存在することが知られている．MSA（Multiple Sequence Alignment）は，こうした類縁タンパク

☆2 https://predictioncenter.org/casp14/

質のアミノ酸配列の集合を，アミノ酸単位で対応づけて整列したものである（図 -2）．これにより，タンパク質上の各位置における，進化過程でのアミノ酸種ごとの出現傾向を知ることができる．この出現傾向は，たとえば溶媒と接するタンパク質表面には親水性アミノ酸が好まれる，などといった立体構造上の制約を反映していると考えられ，立体構造予測にとって大きな情報を与え得る．実際，これまでにも MSA の利用により，二次構造と呼ばれるタンパク質の部分的な規則構造の予測精度向上がもたらされている．現在，さまざまな生物種のゲノム解読が進んでおり，解読されたゲノムにコードされているタンパク質のアミノ酸配列の情報を用いて，任意の対象タンパク質に対して，大量の類縁タンパク質からなる MSA を計算することが，多くの場合可能である．

鋳型構造

先述した共通祖先タンパク質から派生してきたと考えられる類縁タンパク質は，その立体構造も相互に似ている．また多様なタンパク質の構造データの蓄積に伴い，共通祖先から派生してきたものか否か明瞭ではないタンパク質同士の間にも立体構造の類似性が認められる例も少なからず見つかる一方，新規立体構造「発見」の割合は年々減少している．こうした観測を背景に，現在では，自然界に存在する大部分のタンパク質の立体構造パターンはかなり限られるという見方が支配的になっている．実際，既知立体構造を鋳型として用いた構造未知タンパク質のモデリング（異なる種類のアミノ酸に共通する主鎖構造を既知立体構造から借用し，アミノ酸種に固有の側鎖構造を対象のものに変換する）も多くの場面で有効なものとなっている．

AlphaFold

AlphaFold は，その名称から連想される通り，DeepMind 社によって開発された手法である．AlphaFold はその前処理の段階で，予測対象タンパク質に関する MSA の計算と鋳型となり得る構造の検索を行い，それらの情報を入力として利用している．ここで読者は，なぜこの手法が画期的と言えるのか，疑問に思うかもしれない．この問いに対する答えは，最終章で紹介する．

AlphaFold 本体部分のネットワークは，前後半の 2 つに大きく分けられる．前半部分は Evoformer と呼ばれる同等の 48 ブロックから成り，対象タンパク質は，MSA を構成するタンパク質数 s ×対象タンパク質の配列長 r × c（チャンネル数 =256）サイズのテンソル（MSA representation；以降 MSA 表現）および r × r × c（チャンネル数 =128）のサイズのテンソル（pair representation；ペア表現）で表現される．ここでペア表現は，立体構造上での任意の 2 つの位置の近接性を表すのに利用される．要素の初期値は 0 で，前処理の段階で同定された鋳型構造の情報から計算される特徴量が attention network によって変換され，ペア表現に追加される．各ブロックでは，MSA 表現とペア表現が相互に影響を及ぼしつつ複数のレイヤで処理され，各レイヤ

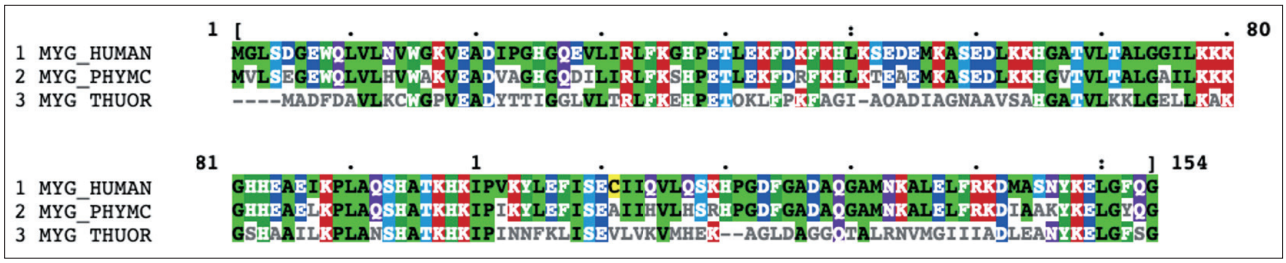

図 -2　ミオグロビンの MSA の例．上段）ヒト，中段）マッコウクジラ，下段）クロマグロのミオグロビンのアミノ酸配列

の出力が残差接続により各々の表現に追加され，両表現が更新される．これを次のブロックの入力とし，合計 48 ブロックの計算を通して，それら表現が改善される．後半部分では，前半で改善されたペア表現および MSA 表現の 1 行目の線形射影により求められる 1 × r × c（チャンネル数 =384）のテンソルを入力として，各アミノ酸の原子座標情報に相当する（原点からの）並進と回転ベクトルの組みが計算される．後半部分およびネットワーク全体の双方で，実際の立体構造と出力（座標）との差を反映させた損失関数を用いてモデルの学習を行っている．

予測精度とその効果

これまでの予測法は，既知立体構造を鋳型として用いたモデリング，あるいは MSA を利用したアミノ酸ペアの立体構造上での近接性予測のいずれかに大きく依拠したものがほとんどであった．両アプローチを統合し，かつ本格的に attention を利用したのはおそらく AlphaFold 2（2018 年に開催された CASP13 で発表された AlphaFold はアミノ酸ペアの立体構造上での距離予測に基づいていた）が初めてである．従来法では，タンパク質のコア領域と呼ばれる類縁タンパク質で保存された領域以外についての予測精度に問題があったが，AlphaFold ではこの点も克服されているようである．AlphaFold は大量配列データと既知立体構造情報を利用した深層学習モデルであり，従来法に比べ，帰納的予測の及ぶ範囲を格段に拡大したと言える．これは同時に，大部分のタンパク質の立体構造パターンはかなり限られるという従来の仮説が妥当であることも示唆している．

AlphaFold の出現により，多くのタンパク質について非常に正確な立体構造モデルが得られるようになった（DeepMind 社はそのデータベースを公開している☆3)）．これにより，タンパク質間あるいはタンパク質と薬剤候補分子などの相互作用に関する研究が一層促進されるものと考えられる．ただし，AlphaFold であってもすべてのタンパク質について正確な立体構造を予測できるという訳ではない．今後，実験あるいは計算機による立体構造決定や予測は，AlphaFold により正確な立体構造モデルが得られないタンパク質中心に行われていくようになるのかもしれない．現在の生物学では，大量情報が利用可能な分野が多くあり，AlphaFold の成功に刺激を受け，今後タンパク質以外の分野でも深層学習の利用が一層活発になるであろう．

（2021 年 9 月 7 日受付）

☆3 https://alphafold.ebi.ac.uk

富井健太郎
k-tomii@aist.go.jp
1998 年京都大学大学院理学研究科博士後期課程生物科学専攻修了．博士（理学）．生物分子工学研究所（BERI）および UC Berkeley ポスドクを経て，2001 年産業技術総合研究所研究員，2008 年同主任研究員，2012 年同研究チーム長．