

AI 特許紹介(32)

AI 特許を学ぶ！究める！

～Evoformer を実装した AlphaFold2 特許～

2021 年 9 月 10 日

河野特許事務所

所長弁理士 河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第 4 次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要

特許出願人 DeepMind

出願日 2020 年 12 月 1 日

公開日 2021 年 6 月 3 日

公開番号 US2021/0166779

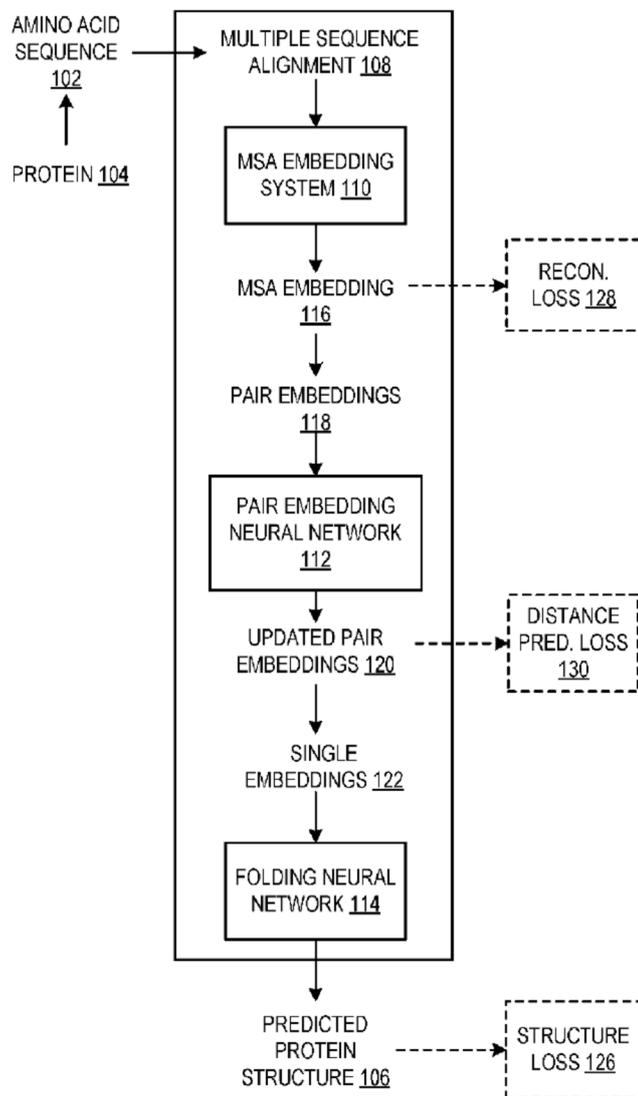
発明の名称 セルフアテンションニューラルネットワークを使用したアミノ酸配列からのタンパク質構造予測

779 特許は、アミノ酸配列からタンパク質の折り畳み構造を予測する Alphafold2 特許に関する。Alphafold2 は、タンパク質構造を予測する CASP13(Critical Assessment of Protein Structure Prediction)コンペティションで第 1 位となった DeepMind 社の AlphaFold の進化版であり、精度向上のためにアテンション機構を有する Evoformer を採用している。本稿では Evoformer の一部に利用されているグリッドトランスフォーマー機構を解説するとともに、Nature に投稿された論文の一部について解説する。

2.特許内容の説明

図1は、タンパク質構造予測システム100を示す。

PROTEIN STRUCTURE PREDICTION SYSTEM 100



構造予測システム100は、タンパク質104のアミノ酸配列102を定義するデータを処理し、タンパク質104の予測構造106を生成する。アミノ酸配列102の各アミノ酸は、アミノ官能基およびカルボキシル官能基、ならびにアミノ酸に特異的な側鎖（すなわち、原子の基）を含む有機化合物である。

予測構造106は、タンパク質104がタンパク質フォールディングを受けた後のタンパク質104のアミノ酸配列102内の原子の三次元（3D）構成の推定値を定義する。

アミノ酸配列102は、任意の適切な数値形式で表すことができる。例えば、アミノ酸配列102は、ワンホットベクトルの配列として表すことができる。タンパク質104の

予測構造 106 は、構造パラメータのセットの値によって定義される。構造パラメータのセットは、タンパク質 104 中の各アミノ酸について、(i)位置パラメータ、および(ii)回転パラメータを含む。

予測構造 106 を生成するために、構造予測システム 100 は、タンパク質 104 のアミノ酸配列 102 に対応するマルチプルシーケンスアラインメント (MSA) 108 を取得する。

MSA108 は、アミノ酸配列 102 と、例えば、他のもの (例えば相同タンパク質) からの複数の追加のアミノ酸配列との配列アラインメントを特定する。MSA108 は、例えば、任意の適切な計算配列アラインメント技術、例えば、プログレッシブアラインメント構築を使用してアミノ酸配列のデータベースを処理することによって生成される。

構造予測システム 100 は、(1)マルチプルアラインメント埋め込みシステム 110 (すなわち、「MSA 埋め込みシステム」)、(2)ペア埋め込みニューラルネットワーク 112、および、(3)フォールディングニューラルネットワーク 114 を使用して、アミノ酸配列 102 および MSA108 から予測構造 106 を生成する。「埋め込み embedding」は、数値の順序付けられた集合、例えば、数値のベクトルまたは行列である。

MSA 埋め込みシステム 110 は、MSA108 を処理して MSA 埋め込み 116 を生成する。MSA 埋め込み 116 は、MSA108 の各アミノ酸配列中の各アミノ酸に対応するそれぞれの埋め込みを含む MSA108 の代替表現である。

MSA 埋め込みは、次元数 $M \times N \times E$ を有するマトリックスとして表すことができる。ここで、 M は、MSA108 の配列の数であり、 N は、MSA108 の各アミノ酸配列のアミノ酸の数 (これは、アミノ酸配列の長さと同じ)、 E は MSA108 の各アミノ酸に対応する埋め込みの次元数である。

一例では、MSA 埋め込みシステム 110 は、MSA 埋め込み M を以下のように生成することができる。

$$M = h_{\theta}(\mathbb{I}_{AA}) \oplus h_{\phi}(\mathbb{I}_{MSA}) \quad (1)$$

ここで、 \mathbb{I}_{AA} は、ワンホットアミノ酸埋め込みベクトルの 1-D アレイとしてのアミノ酸配列 102 の表現であり、 $h_{\theta}(\cdot)$ は、1 次元配列 \mathbb{I}_{AA} の各埋め込みに適用される学習済み線形射影演算であり、 \mathbb{I}_{MSA} は、ワンホットアミノ酸埋め込みベクターの 2D アレイとしての MSA 108 の表現であり (2D アレイの各行は MSA のそれぞれのアミ

ノ酸配列に対応する)、 $h\phi(\cdot)$ は、2次元配列 IIMSA の各埋め込みに適用される学習済み線形射影演算であり、 \otimes 演算は、 $h\phi$ (IIMSA) の各行に $h\theta$ (IIAA) を追加することを意味する。

MSA 埋め込み 116 を生成した後、構造予測システム 100 は、MSA 埋め込み 116 を、ペア埋め込み 118 の集合としての代替表現に変換する。埋め込み 118 の各対は、タンパク質のアミノ酸配列 102 内のアミノ酸の対に対応する。アミノ酸配列 102 のアミノ酸の対は、アミノ酸配列 102 の第 1 のアミノ酸および第 2 のアミノ酸を指し、アミノ酸の対のセットは、以下によって与えられる。

$$\{(A_i, A_j): 1 \leq i, j \leq N\} \quad (2)$$

ここで、 N はアミノ酸配列 102 のアミノ酸の数、すなわちアミノ酸配列 102 の長さであり、 A_i はアミノ酸配列 102 の位置 i のアミノ酸であり、 A_j は配列における位置 j のアミノ酸である。

ペア埋め込み 118 の集合は、次元 $N \times N \times F$ を有する行列として表すことができる。ここで、 N は、アミノ酸配列 102 内のアミノ酸の数であり、 F は、各ペア埋め込み 118 の次元である。

構造予測システム 100 は、例えば、MSA 埋め込み 116 の外積をそれ自体で計算し、外積の MSA 寸法をマージナライズ (例えば、平均化) し、そして外積の埋め込み寸法を結合することによって、MSA 埋め込み 116 からペア埋め込み 118 を生成する。例えば、ペア埋め込み 118 は、以下のように生成される。

$$L_{mic} = M_{mlo} \overline{L}_{oc} \quad (3)$$

$$R_{mrd} = M_{mro} \overline{R}_{od} \quad (4)$$

$$P_{lre} = L_{mic} R_{mrd} \overline{C}_{cde} \quad (5)$$

ここで、 M は、MSA 埋め込み 116 を示し (例えば、MSA 内の配列を索引付けする第 1 の次元、MSA 内の各配列内のアミノ酸を索引付けする第 2 の次元、および各アミノ酸の埋め込みのチャンネルを索引付けする第 3 の次元を有する)、 L は、 M を乗算して「左」にエンコードされた埋め込み L を生成する重み行列を示し、 R は、 M を乗算して「右」にエンコードされた埋め込み R を生成する重み行列を示し、 C は、左エンコードと右エンコードの埋め込みを組み合わせてペア埋め込み P を生成する重み行列を示

す。式 (3) ~ (5) は、アインシュタインの縮約記法を使用し、ここで、方程式の右側にのみ存在する繰り返されるインデックスは、暗黙的に合計される。

換言すれば、構造予測システム 100 は、MSA 埋め込み 116 の「外積平均」を計算することによって、MSA 埋め込み 116 からペア埋め込み 118 を生成することができ、ここで、MSA 埋め込みは、埋め込みの $M \times N$ アレイとして見られる（配列の各埋め込みは次元 E を持つ）。外積平均は、MSA 埋め込み 116 に適用されると、ペア埋め込み 118 を定義する埋め込みの $N \times N$ 配列（すなわち、 N はタンパク質中のアミノ酸の数）を生成する一連の操作を定義する。

MSA 埋め込み 116 の外積平均を計算するために、システム 100 は、外積平均操作を MSA 埋め込み 116 に適用し、外積平均操作の結果としてペア埋め込み 118 を識別することができる。外積平均を計算するために、システムはテンソル $A(\cdot)$ を生成し、たとえば、次の式で与えられる。

$$A(res1, res2, ch1, ch2) =$$

$$\frac{1}{M} \sum_{m=1 \dots M} LeftAct(m, res1, ch1) \cdot RightAct(m, res2, ch2)$$

$$res1, res2 \in \{1, \dots, N\}$$

ここで、 N はタンパク質内のアミノ酸の数、 $ch1, ch2 \in \{1, \dots, E\}$ 、ここで、 E は、MSA 埋め込みを表す埋め込みの $M \times N$ 配列の各埋め込みのチャンネル数、 M は、MSA 埋め込みを表す埋め込みの $M \times N$ 配列の行数、 $LeftAct(m, res1, ch1)$ は、MSA 埋め込みを表す埋め込みの $M \times N$ 配列で、「 m 」でインデックス付けされた行と「 $res1$ 」でインデックス付けされた列にある埋め込みのチャンネル $ch1$ に適用される線形演算（たとえば、行列乗算によって定義される）、 $RightAct(m, res2, ch2)$ は、MSA 埋め込みを表す埋め込みの $M \times N$ 配列で、「 m 」でインデックス付けされた行と「 $res2$ 」でインデックス付けされた列にある埋め込みのチャンネル $ch2$ に適用される線形演算（たとえば、行列乗算によって定義される）である。

外積平均の結果は、テンソル A の $(ch1, ch2)$ 次元を平坦化し、線形に投影することによって生成される。ペア埋め込みニューラルネットワーク 112 は、ペア埋め込みのコレクションを処理して、ペア埋め込みの値を更新する、すなわち、更新されたペア埋め込み 120 を生成する。ペア埋め込みニューラルネットワーク 112 は、1 つまたは複数のセルフアテンションニューラルネットワーク層を使用してそれらを処理すること

によってペア埋め込みを更新する。

セルフアテンション層は一般に、埋め込みのコレクションを更新するニューラルネットワーク層を指す。つまり、埋め込みのコレクションを受信し、更新された埋め込みを出力する。特定の埋め込みを更新するために、セルフアテンション層は、特定の埋め込みと1つ以上の選択された埋め込みのそれぞれとの間のそれぞれの「アテンションの重み」を決定し、次に、(i) アテンションの重み、および (ii) 選択した埋め込みを使用して特定の埋め込みを更新する。セルフアテンション層は、選択された埋め込み「上 over」にアテンションを使用して、特定の埋め込みを更新する。

一例では、セルフアテンション層は、入力埋め込みのコレクションを受け取り、埋め込み x_i を更新するために、セルフアテンション層は、アテンションの重み $[a_j]_{j=1}^N$ を決定することができる。ここで、 $S_i \subset \{1, \dots, N\}$ および α_j は、次のように、 x_i と x_j の間のアテンションの重みを示す。

$$[a_j]_{j \in S_i} = \text{softmax} \left(\frac{(W_q x_i) K_{S_i}^T}{c} \right) \quad (6)$$

$$K_{S_i} = [W_k x_j]_{j \in S_i} \quad (7)$$

ここで、 W_q と W_k は学習されたパラメータ行列であり、 $\text{softmax}(\cdot)$ はソフトマックス正規化演算を示し、 c は定数である。アテンションの重みを使用して、セルフアテンション層は埋め込み x_i を次のように更新する。

$$x_i \leftarrow \sum_{j \in S_i} a_j \cdot (W_v x_j) \quad (8)$$

ここで、 W_v は学習されたパラメータ行列である。

セルフアテンションニューラルネットワーク層に加えて、ペア埋め込みニューラルネットワークは、セルフアテンション層、例えば線形ニューラルネットワーク層とインターリーブされ得る他のニューラルネットワーク層を含む。一例では、ペア埋め込みニューラルネットワーク 112 は、「グリッドトランスフォーマーニューラルネットワークアーキテクチャ」と呼ばれるニューラルネットワークアーキテクチャを有する（後述する図 2 参照）。

構造予測システム 100 は、ペア埋め込み 118 を使用して、タンパク質 104 内の各アミノ酸に対応するそれぞれの「単一」埋め込み 122 を生成する。一例では、構造予測シ

システムは、アミノ酸 i に対応する単一埋め込み S_i を以下のように生成することができる。

$$S_i = \sum_{j=1}^N P_{i,j} \quad (9)$$

ここで、 $P_{i,j}$ はアミノ酸ペア (A_i, A_j) に対応する埋め込みペア、 A_i はアミノ酸配列 102 の位置 i のアミノ酸、 A_j はアミノ酸配列 102 の位置 j のアミノ酸である。

フォールディングニューラルネットワーク 114 は、単一の埋め込み 122、ペアの埋め込み 120、またはその両方を含む入力を処理することによって、タンパク質の予測される構造 106 を定義する構造パラメータの値を指定する出力を生成するよう構成される。

一例では、フォールディングニューラルネットワーク 114 は、1 つまたは複数のニューラルネットワーク層、例えば、畳み込みまたは完全結合ニューラルネットワーク層を使用して、単一の埋め込み 122、ペアの埋め込み 120、またはその両方を処理して、タンパク質 104 の構造パラメータの値を出力する。

図 2 は、グリッドトランスフォーマーニューラルネットワークアーキテクチャ 200 を示す。

GRID TRANSFORMER NEURAL NETWORK ARCHITECTURE 200

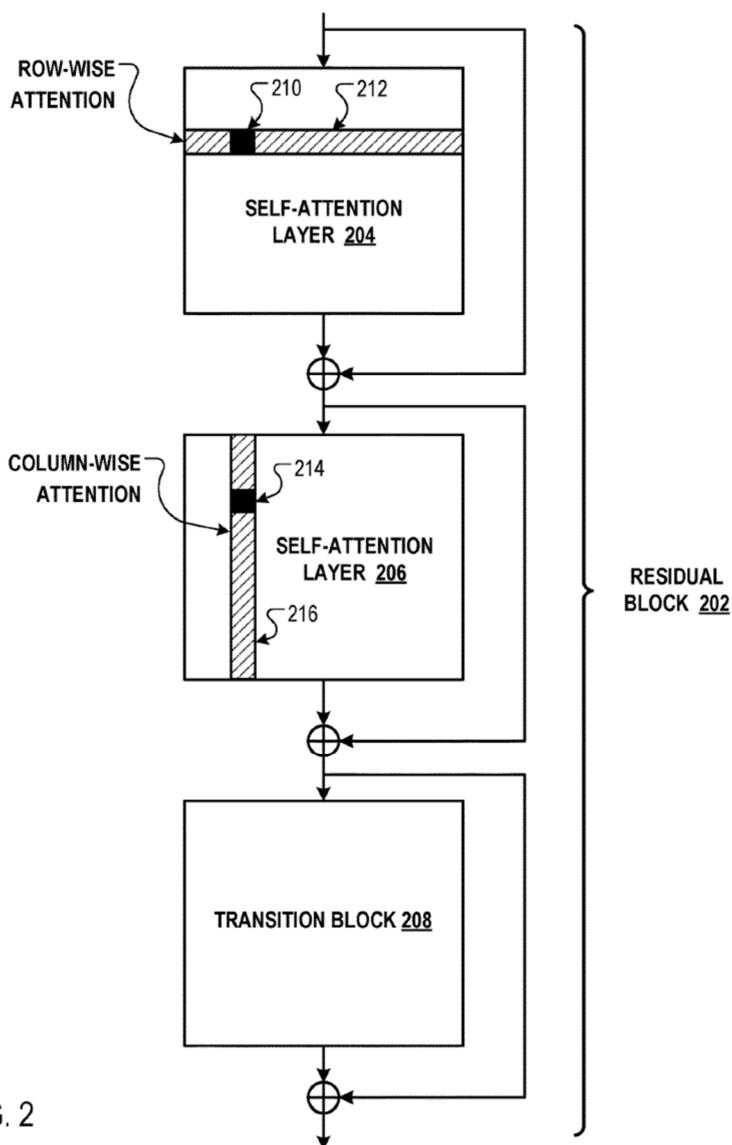


FIG. 2

グリッドトランスフォーマーニューラルネットワークは、初期値を有する埋め込みのコレクションを受信し、1つまたは複数の残余ブロック 202 を介して埋め込みのコレクションを処理し、更新された埋め込みのコレクションの値を出力する。

グリッドトランスフォーマーアーキテクチャの各残余ブロック 202 は、「行方向のセルフアテンション」を実装するセルフアテンション層 204、「列方向のセルフアテンション」を実装するセルフアテンション層 206、および遷移ブロック 208 を含む。

行方向のセルフアテンション層 204 は、埋め込みの 2 次元配列内の埋め込みと同じ行にある埋め込みのみにアテンションを使用して、各埋め込みを更新する。

例えば、行方向のセルフアテンション層 204 は、同じ行の埋め込み（ハッチングされた領域 212）に対するアテンションを使用して、埋め込み 210（黒四角）を更新する。残余ブロック 202 はまた、行方向のセルフアテンション層 204 への入力を、行方向のセルフアテンション層 204 の出力に追加する。

列方向のセルフアテンション層 206 は、埋め込みの 2 次元配列内の埋め込みと同じ列にある埋め込みのみにアテンションを使用して、各埋め込みを更新する。例えば、列方向のセルフアテンション層 206 は、同じ列内の埋め込み（ハッチングされた領域 216）に対するセルフアテンションを使用して、埋め込み 210（黒四角）を更新する。残余ブロック 202 はまた、列方向のセルフアテンション層への入力を、列方向のセルフアテンション層の出力に追加する。

遷移ブロック 208 は、1 つまたは複数の線形ニューラルネットワーク層を使用してそれら进行处理することによって、埋め込みを更新する。残余ブロック 202 はまた、遷移ブロックへの入力を遷移ブロックの出力に追加する。

3. クレーム

779 特許のクレーム 1-7 は以下の通りである。

1. アミノ酸配列によって指定されるタンパク質の予測される構造を決定するために、1 つまたは複数のデータ処理装置によって実行される方法において、
タンパク質のマルチプルアラインメント（MSA:multiple sequence alignment）を取得し、
マルチプルアラインメントから、タンパク質のアミノ酸配列内のアミノ酸の各ペアについて、アミノ酸のペアのそれぞれの初期埋め込みを決定し、
アミノ酸の各ペアの最終埋め込みを生成すべく、複数のセルフアテンションニューラルネットワーク層を含むペア埋め込みニューラルネットワークを使用してアミノ酸のペアの初期埋め込みを処理し、
アミノ酸の各ペアの最終埋め込みに基づいて、タンパク質の予測される構造を決定する。
2. クレーム 1 の方法において、ペア埋め込みニューラルネットワークの各セルフアテンションニューラルネットワーク層は、下記操作を実行するように構成されている：

アミノ酸の各ペアの現在の埋め込みを受け取り、
アミノ酸のペアの現在の埋め込みに対するアテンションを使用して、アミノ酸の各ペアの現在の埋め込みを更新する。

3. クレーム2の方法において、
アミノ酸の各ペアについて、アミノ酸のペアの現在の埋め込みに対するアテンションを使用してアミノ酸のペアの現在の埋め込みを更新することは、以下を含む：
アミノ酸のペアの現在の埋め込みの適切なサブセットのみに対するアテンションを使用して、アミノ酸のペアの現在の埋め込みを更新する。

4. クレーム3の方法において、
1つまたは複数のセルフアテンションニューラルネットワーク層は、行ごとのセルフアテンションニューラルネットワーク層であり、
行ごとのセルフアテンションニューラルネットワーク層ごとに、アミノ酸のペアの現在の埋め込みの適切なサブセットのみに対してアテンションを使用して、アミノ酸のペアの現在の埋め込みを更新することは、以下を含み
アミノ酸のペアの現在の埋め込みの配置におけるアミノ酸のペアの現在の埋め込みと、二次元配列にて同じ行に位置するアミノ酸のペアの現在の埋め込みのみに対してアテンションを使用して、アミノ酸のペアの現在の埋め込みを更新する。

5. クレーム4の方法において、
1つまたは複数のセルフアテンションニューラルネットワーク層は、列ごとのセルフアテンションニューラルネットワーク層であり、
列ごとのセルフアテンションニューラルネットワーク層ごとに、アミノ酸のペアの現在の埋め込みの適切なサブセットのみに対しアテンションを使用して、アミノ酸のペアの現在の埋め込みを更新することは、以下を含み、
アミノ酸のペアの現在の埋め込みの配置におけるアミノ酸のペアの現在の埋め込みと、二次元配列にて同じ列にあるアミノ酸のペアの現在の埋め込みのみに対しアテンションを使用して、アミノ酸のペアの現在の埋め込みを更新する。

6. クレーム5の方法において、
ペア埋め込みニューラルネットワークの複数のセルフアテンションニューラルネットワーク層は、行ごとのセルフアテンションニューラルネットワーク層と列ごとのセルフアテンションニューラルネットワーク層の交互のシーケンスを含む。

7. クレーム1の方法において、

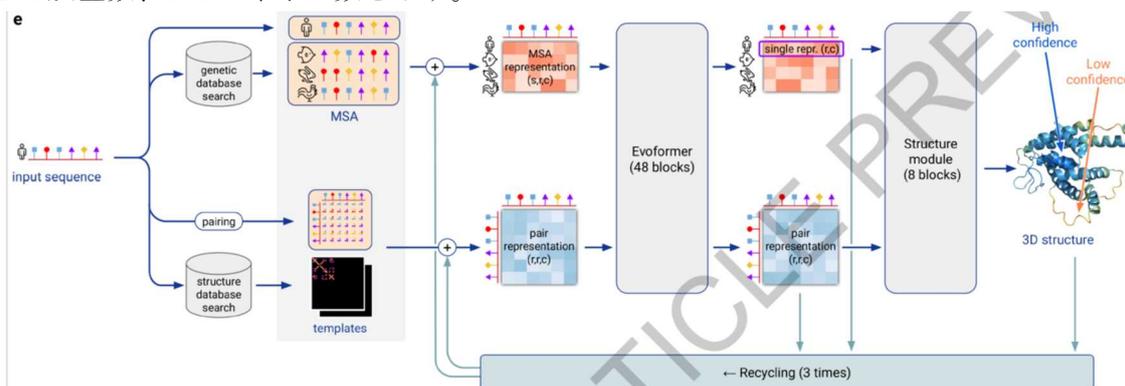
ここで、アミノ酸の各ペアの最終的な埋め込みに基づいてタンパク質の予測される構造を決定することは、以下を含む：

アミノ酸のペアの最終的な埋め込みに基づいて、タンパク質のアミノ酸配列における各アミノ酸のそれぞれの最初の埋め込みを決定し、アミノ酸配列への各アミノ酸の最初の埋め込みに基づいて、タンパク質の予測される構造を決定する。

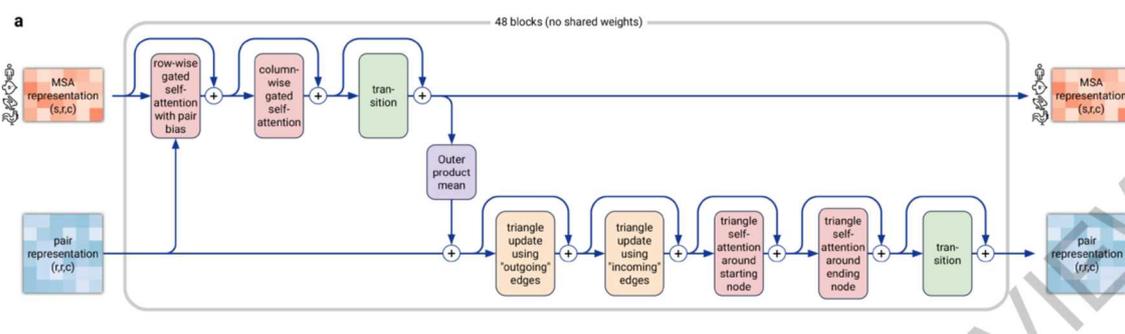
4. 本特許に関する論文

本特許に関連する論文¹「Highly accurate protein structure prediction with AlphaFold」が John Jumper 氏らにより Nature 誌に発表されている。

下記図 1e は、AlphaFold2 のモデルアーキテクチャを示す。なお、 s はシーケンス数、 r は残基数、 c はチャンネル数を示す。



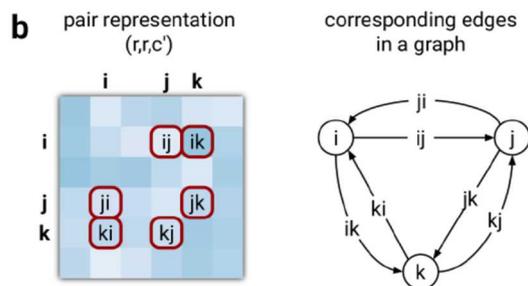
下記図 3a は Evoformer のブロック図である。



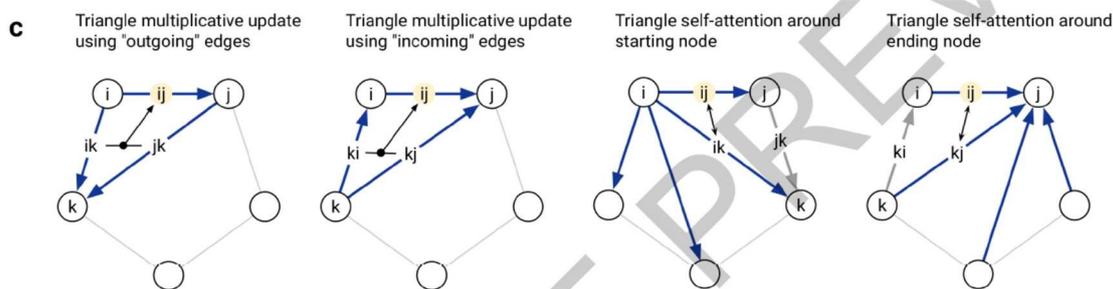
¹John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli & Demis Hassabis 著 “Highly accurate protein structure prediction with AlphaFold” <https://doi.org/10.1038/s41586-021-03819-2>

Evoformer の前段部分に本特許のグリッドトランスフォーマーアーキテクチャが設けられている。さらに、後段部分には、トライアングル乗法更新及びトライアングルセルフアテンションアーキテクチャが設けられている。

下記図 3b は、グラフ内の有向エッジとして解釈されるペア表現である。ペア表現の要素は、残基間の関係に関する情報をエンコードする。



アミノ酸のペアワイズ記述を単一の 3D 構造として表現できるようにするには、距離の三角不等式を含む多くの制約を満たす必要がある。この直感に基づいて、3つの異なるノードを含むエッジの三角形の観点からペア表現の更新操作を調整する。下記図 3c は、トライアングル乗法更新とトライアングルセルフアテンションを示す。



円は残基を示し、ペア表現のエントリは有向エッジとして示され、各図で更新されるエッジは ij である。特に、三角形の「欠落エッジ missing edge」を含めるために軸方向のアテンションにロジットバイアスを追加し、2つのエッジを使用して欠落している3番目のエッジを更新する非アテンション更新操作「三角形乗法更新」を定義する。

三角形乗法更新は、元々、アテンションのより対称的でやや安価な代替として開発された。アテンションまたは乗法更新のみを使用するネットワークは、どちらも高精度の構造を生成できる。ただし、2つの更新の組み合わせにより精度を向上させることができる。

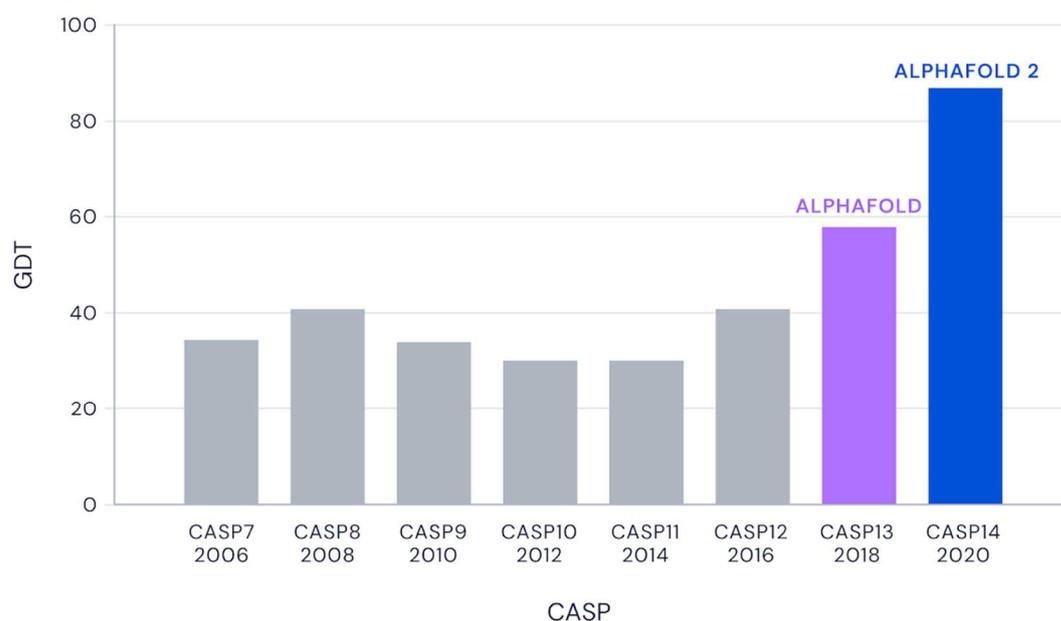
また、MSA 表現内で軸方向アテンションの変形を使用する。MSA でのシーケンスごとのアテンション中に、ペアスタックから追加のロジットを投影して MSA アテンショ

ンにバイアスをかける。これにより、ペア表現から MSA 表現に戻る情報フローが提供され、ループが閉じられ、Evoformer ブロック全体がペア表現と MSA 表現の間で情報を完全に混合し、構造モジュール内での構造生成の準備ができる。

下記図は CASP における GDT (Global Distance Test) スコアを示すグラフである²。スコアが 90 を超えれば十分とされるところ、DeepMind 社の Alphafold は 2 回目の出場でこの目標に到達した。

本稿で述べたように様々な分野で Transformer が活用され、精度向上が図られている。

Median Free-Modelling Accuracy



以上

著者紹介

河野英仁

河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コ

² DeepMind 社 HP より 2020 年 12 月 3 日
<https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology>

ース修了。

[AI 特許コンサルティング](#)、[医療 AI 特許コンサルティング](#)の他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウェア特許(共著)」、「FinTech 特許入門」、「[AI/IoT 特許入門 2.0](#)」、「[ブロックチェーン 3.0\(共著\)](#)」がある。