AI 特許紹介(15)
～AlphaFold 特許～

2020 年 5 月 20 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。

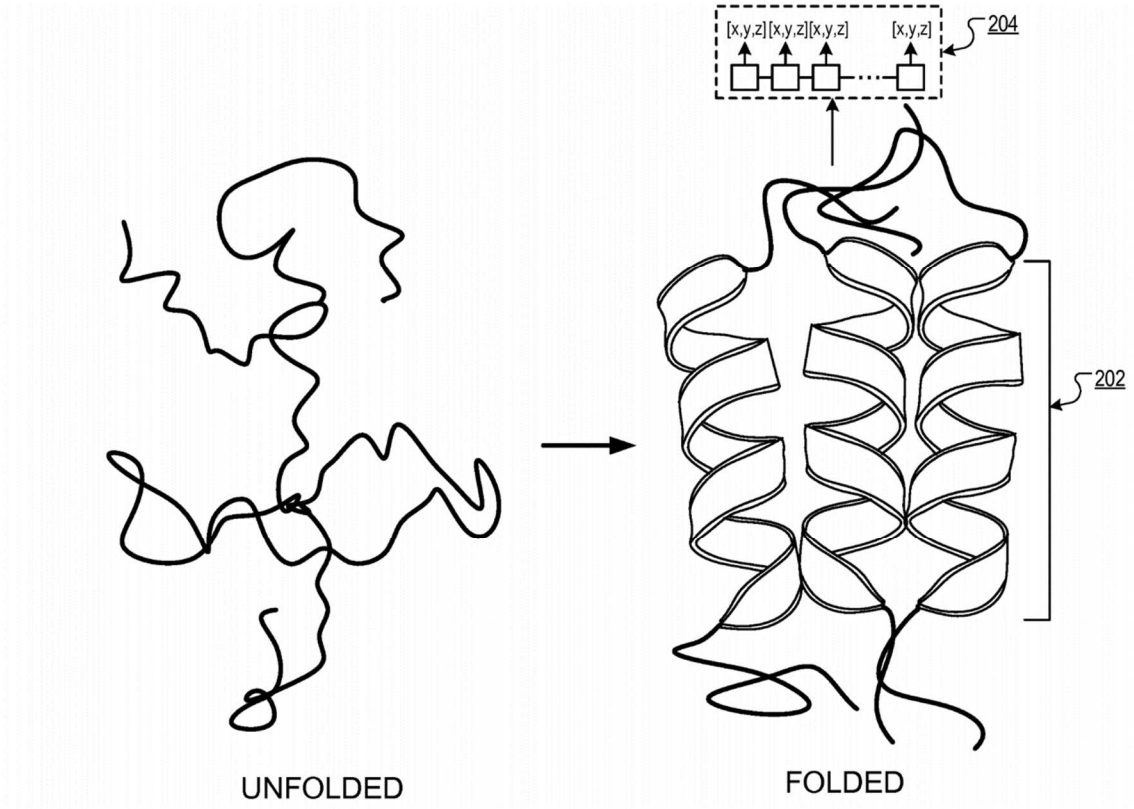
本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要
特許出願人　DeepMind Technologies Limited
出願日　2019 年 9 月 16 日
公開日　2020 年 3 月 26 日
公開番号 WO2020058176
発明の名称　タンパク質構造を決定するための機械学習

176 特許は、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 に関する特許である。176 特許はアミノ酸のシーケンスからタンパク質の折りたたみ後の 3 次元構造を予測する。本稿は AlphaFold の AI アルゴリズムを 176 特許に基づき、解説するとともに、Nature 誌に掲載された論文を紹介する。

2.特許内容の説明
タンパク質は、アミノ酸のシーケンス(配列)からなる。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、原子団）を含む有機化合物である。タンパク質の折りたたみは、アミノ酸のシーケンスが 3 次元構成に折りたたまれる物理的なプロセスを指す。下記図は折りたたまれていないタンパク質

と折りたたまれたタンパク質を示す説明図である。



折りたたまれたタンパク質の構造は、一連の構造パラメータの値によって定義できる。たとえば、204 で示されているように、構造パラメータは、折りたたまれたタンパク質のアミノ酸のバックボーン原子のそれぞれの位置を表す 3D 数値座標（たとえば、[x、y、z /座標]）のシーケンスである。

下記図 6 は、距離予測システム 528 の処理を示す説明図である。176 特許では機械学習によりタンパク質の距離マップ 608 を生成する。

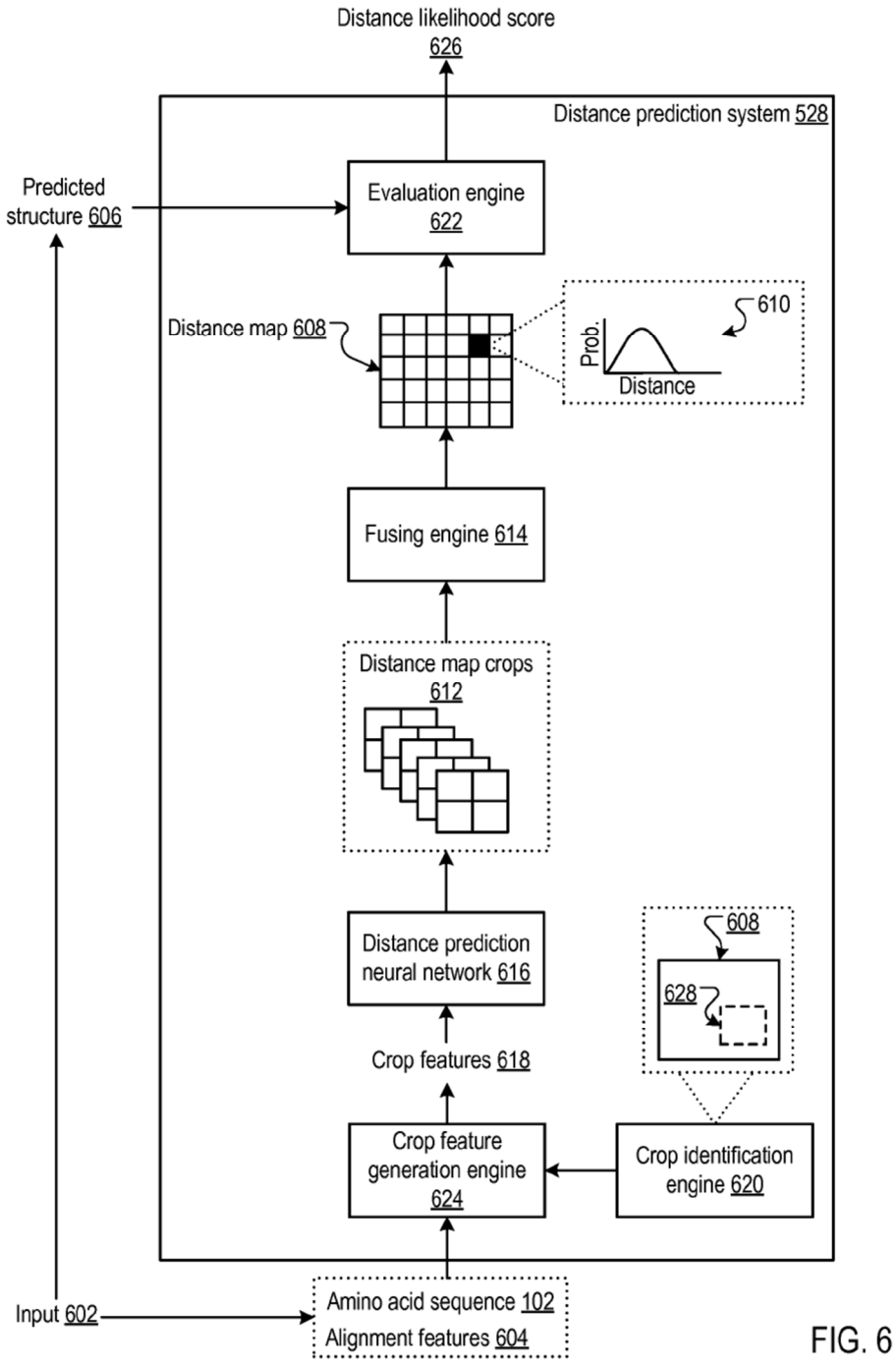


FIG. 6

距離予測システム 528 は、アミノ酸シーケンス 102、アラインメント特徴 604、およびアミノ酸シーケンス 102 の予測構造 606 の表現を含む入力 602 を受け取る。システム 528 は、入力 602 を処理して距離尤度スコア 626 を生成するように構成される。

距離尤度スコア 626 は、(i)予測構造 606 のアミノ酸シーケンス 102 におけるアミノ酸のペア間の距離と、(ii)アミノ酸シーケンス 102 の実際の構造におけるアミノ酸シーケ

ンス 102 のアミノ酸のペア間の推定距離との間の差に基づいて、予測構造 606 の尤度を定義する。

距離尤度スコア 626 を生成するために、システム 528 は、アミノ酸シーケンス 102 の実際の（例えば、実験的に決定された）構造におけるアミノ酸シーケンス 102 内のアミノ酸の各ペア間の推定距離を特徴付ける距離マップ 608 を生成する。

距離マップ 608 を生成するために、システム 528 は、一組の距離マップクロップ 612 を生成する。各距離マップクロップ 612 は、完全距離マップ 608 の適切なサブセットの推定である。具体的には、各距離マップクロップ 612 は、(i) アミノ酸シーケンス 102 の 1 つまたは複数の第 1 位置のそれぞれにあるアミノ酸残基と、(ii) アミノ酸シーケンス 102 の 1 つまたは複数の第 2 位置のそれぞれにあるアミノ酸残基との間の予測距離を特徴づける。

第 1 位置は、距離マップ行列 608 のそれぞれの行を特定し、第 2 位置は、距離マップ行列 608 のそれぞれの列を特定する。システム 528 は、融合エンジン 614 を使用して距離マップクロップ 612 を組み合わせ、完全距離マップ 608 を生成する。例えば、複数の距離マップクロップ 612 の平均を完全距離マップ 608 とすることができる。

距離予測ニューラルネットワーク 616 は、拡張畳み込みニューラルネットワーク層、残差ブロック、及び、アテンション層を含み、複数のトレーニング例を含む訓練データセットに基づいて訓練することができる。各トレーニング例には、トレーニングネットワーク入力と、トレーニングネットワーク入力に対応するターゲット距離マップが含まれる。トレーニングネットワークの入力は、既知の構造を持つトレーニングタンパク質から得られる。

評価エンジン 622 は、アミノ酸シーケンス 102 内のアミノ酸の各ペアについて、アミノ酸のペアが予測構造 606 によって定義される距離により分離される距離マップ 608 による確率に基づいて、距離尤度スコア 626 を決定する。

下記図は、距離予測ニューラルネットワークの残差ブロックのアーキテクチャを示す説明図である。

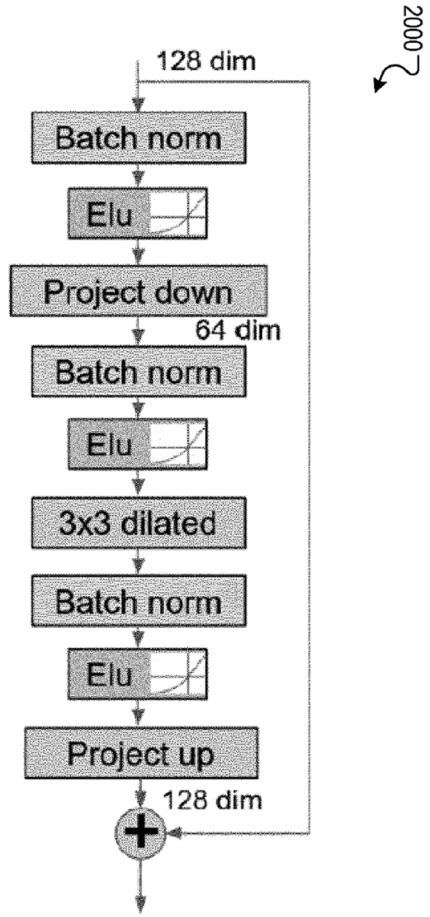


FIG. 20

残差ブロックは、一連のニューラルネットワーク層、3 つの交互配置されるバッチ正規化層、2 つの l x l 投影層、3 x 3 の拡張畳み込み層、および ELU(Exponential Linear Unit)非線形性で構成される。

ブロックの出力は、前のレイヤーの表現に追加される。残差ニューラルネットワークのバイパス接続により、勾配が減少することなくネットワークを通過できるため、非常に深いネットワークのトレーニングが可能になる。

連続する層は、1、2、4、および 8 ピクセルの拡張を循環して、シーケンスおよび MSA(multiple sequence alignment)特徴のトリミングされた領域全体にすばやく情報を伝播できるようにする。本アーキテクチャの距離予測ニューラルネットワークは、220 個のそのような残差ブロックのシーケンスを含み得る。

最後の残差ブロックの後、距離予測ニューラルネットワークは、距離予測ニューラルネットワークによって生成された距離マップクロップの各 i、j コンポーネントに対応

するそれぞれのソフトマックス関数を持つ出力層を含む。

下記図 21 は、タンパク質構造フラグメントを生成するように構成された DRAW 生成ニューラルネットワーク 2100 の例示的なアーキテクチャの図である。

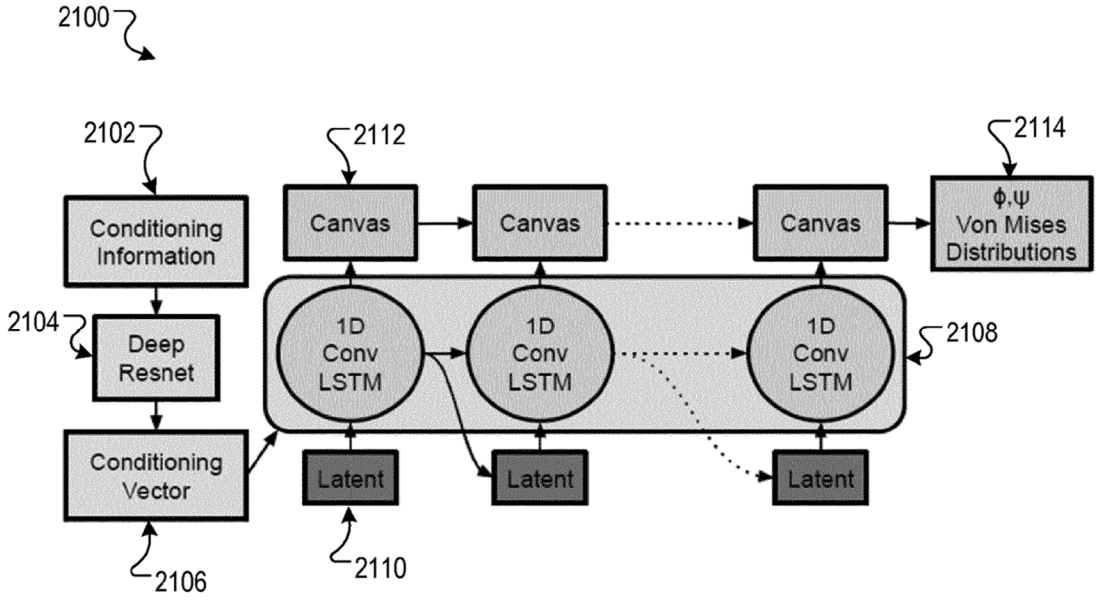


FIG. 21

生成ニューラルネットワーク 2100 は、埋め込みニューラルネットワーク 2104 を使用してアミノ酸シーケンス、及び、アミノ酸シーケンスの MSA 特徴（たとえば、タンパク質のより長いアミノ酸シーケンスのサブシーケンス）を含む 2D 調整情報 2102 を処理して、調整ベクトル 2106 を生成する。

埋め込みニューラルネットワーク 2104 は、1 つ以上の畳み込み残差ブロックを含み、その後に、調整ベクトル 2106 を出力する平均プーリング層が続く。次に、調整ベクトル 2106 は、l-D 畳み込み長期短期記憶（LSTM）畳み込みデコーダサブネットワーク 2108 に渡される。

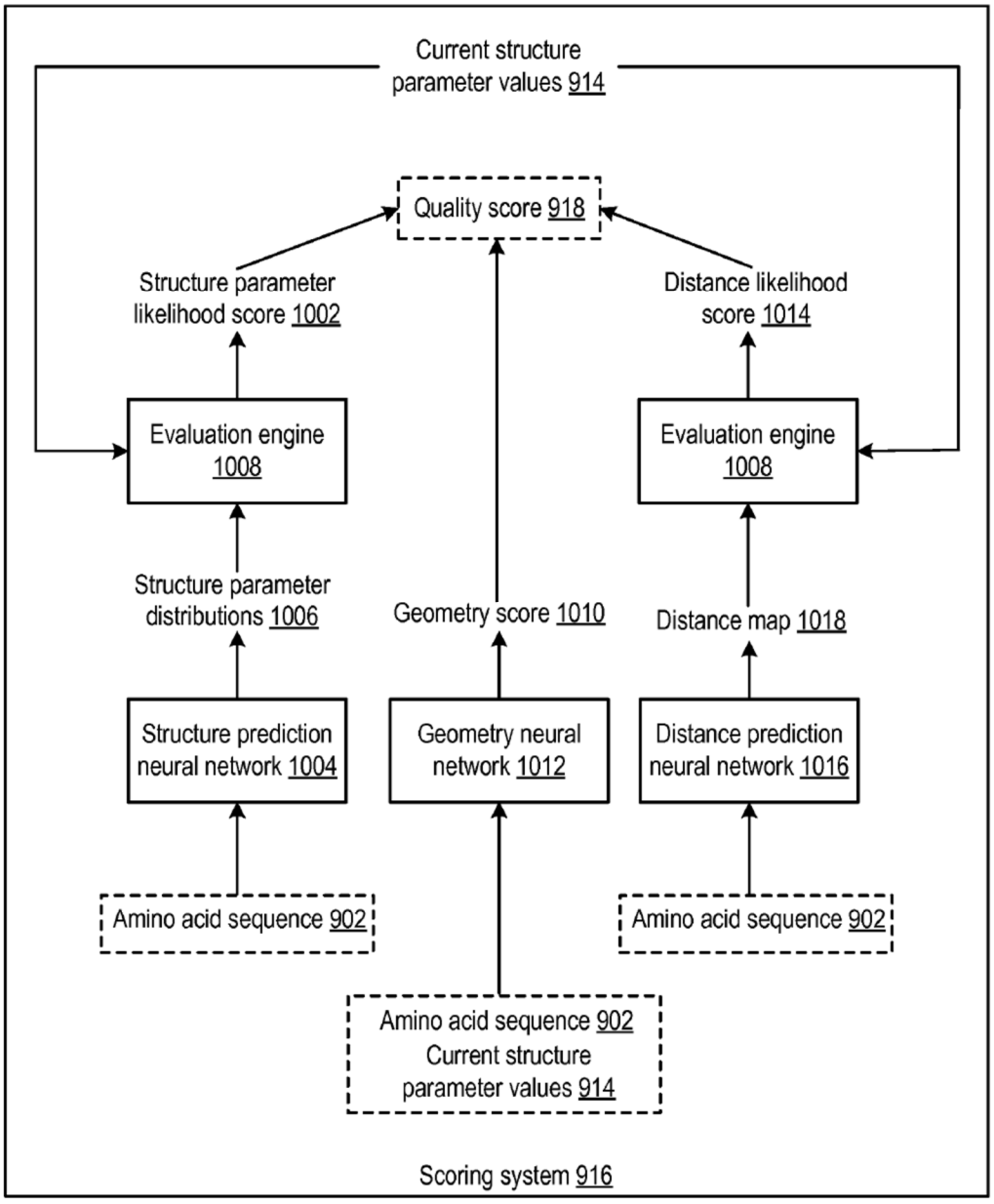
128 個の内部時間ステップのそれぞれで、デコーダサブネットワーク 2108 は、潜在空間にわたる事前確率分布に従って潜在変数 2110 をサンプリングし、潜在変数 2110 および調整ベクトル 2106 を処理して、デコーダサブネットワーク 2108 の内部変数の内部状態を更新する。 なお、事前確率分布は、例えば潜在空間 2110 にわたる標準正規分布であってもよい。

各内部タイムステップで、生成ニューラルネットワーク 2100 は、タイムステップで

のデコーダサブネットワーク 2108 の更新された内部状態を、生成ニューラルネットワーク 2100 のキャンバス内部状態 2112 に追加する。最後の内部時間ステップの後、生成ニューラルネットワーク 2100 のキャンバス内部状態 2112 の値は、タンパク質構造フラグメントの各構造パラメータに対応するそれぞれの確率分布 2114（たとえば、フォンミーゼス分布）を定義する。

その後、任意の所望の数のタンパク質構造フラグメントの構造パラメータ値を、構造パラメータ値にわたる確率分布 2114 に従ってサンプリングすることができる。

176 特許はアミノ酸シーケンス間の距離に基づく距離尤度スコア 626 に加えて、アミノ酸シーケンス間のねじれ角等により定まる構造パラメータ尤度スコア及びジオメトリスコアを最適化し、最終的なタンパク質の予測構造を推測する。下記図 10 は、品質スコアの算出処理手順を示す説明図である。



スコアリングシステム 916 は、最適化システムの更新・反復により、タンパク質のアミノ酸シーケンス 920 の表現および現在の構造パラメータ値 914 を処理し、品質スコア 918 を生成する。品質スコア 918 は、現在の構造パラメータ値 914 によって定義されるタンパク質の予測構造の品質を特徴付ける数値である。

スコアリングシステム 916 は、(i)構造パラメータ尤度スコア 1002、(ii)幾何学スコア 1010、(iii)距離尤度スコア 1018 を生成し、その後それらを組み合わせる（たとえば 、加重線形結合として）品質スコア 918 を生成する。

構造パラメータ尤度スコア 1002 を生成するために、スコアリングシステム 916 は、構造予測ニューラルネットワーク 1004 を使用して、アミノ酸シーケンス 902 の表現を含む入力を処理する。アミノ酸シーケンス 902 の表現は、アミノ酸シーケンス 902 内の各アミノ酸を表すワンホットベクトルのシーケンスである。

アミノ酸シーケンス 902 の表現に加えて、構造予測ニューラルネットワーク 1004 は、例えば、アミノ酸シーケンス 902 を伴う他のタンパク質からのアミノ酸シーケンスの複数シーケンスアラインメント（MSA ：multiple sequence alignment）から得られたデータを含む追加の入力を、処理するよう構成される。

構造予測ニューラルネットワーク 1004 は、構造予測ニューラルネットワーク 1004 の重みに従って構造予測ニューラルネットワークの入力を処理し、各構造パラメータについて、構造パラメータの可能な値にわたるそれぞれの確率分布(構造パラメータ分布 1006)を定義する出力を生成するように構成される。

例えば、構造パラメータがアミノ酸シーケンス 902 のバックボーン原子間のねじれ角のセットである場合、各ねじれ角について、構造予測ニューラルネットワーク 1004 は、下記角度範囲のセットのそれぞれについて各確率を生成することができる。

$$\left[0,\frac{\pi}{3}\right),\left[\frac{\pi}{3},\frac{2\pi}{3}\right),\left[\frac{2\pi}{3},\pi\right),\left[\pi,\frac{4\pi}{3}\right),\left[\frac{4\pi}{3},\frac{5\pi}{3}\right),\left[\frac{5\pi}{3},2\pi\right).$$

構造パラメータ分布 1006 から構造パラメータ尤度スコア 1002 を決定するために、評価エンジン 1008 は、構造予測ニューラルネットワーク 1004 によって生成された構造パラメータの可能な値にわたる対応する確率分布を使用して、各現在の構造パラメータ値 914 の確率を決定する。その後、評価エンジン 1008 は、各現在の構造パラメータ値 914 のそれぞれの確率に基づいて、構造パラメータ尤度スコア 1002 を決定することができる。

さらに幾何学的スコア 1010 を生成するために、スコアリングシステム 916 は、幾何学的ニューラルネットワーク 1012 を使用して、アミノ酸シーケンス 902 および現在の構造パラメータ値 914 の表現を含む入力を処理する。

幾何学的ニューラルネットワーク 1012 は、重みの値に従ってジオメトリニューラルネットワーク入力を処理して、ジオメトリスコア 1010 を生成する。ジオメトリスコア 1010 は、現在の構造パラメータ値 914 によって定義された予測構造とタンパク質 904 の実際の構造との間の類似度の推定値である。例えば、類似度は、現在の構造パラメー

タ値 914 とタンパク質 904 の実際の構造を定義する構造パラメータ値との間の二乗平均平方根偏差であり得る。

スコアリングシステム 916 は、生成した(i)構造パラメータ尤度スコア 1006、(ii)幾何学スコア 1010、(iii)距離尤度スコア 1018 を組み合わせ、品質スコア 918 を生成する。そして構造パラメータ値 914 を調整し、勾配降下法に基づき品質スコアを最適化し、更新反復を行い、構造パラメータの現在値によって定義されるタンパク質の予測構造を決定する。

3.176 特許のクレーム 1,14,15,16,33 及び 39 は以下の通りである。

1. 特定のタンパク質の距離マップを生成するためのコンピュータ実装方法において、

特定のタンパク質は構造内に配置されたアミノ酸残基のシーケンスによって定義され、距離マップは構造内のアミノ酸残基間の推定距離を特徴付けており、

複数の距離マップクロップを生成し、

各距離マップクロップは、タンパク質の構造において、(i) シーケンス内の 1 つ以上の各第 1 位置のそれぞれのアミノ酸残基と、(ii) シーケンス内の 1 つ以上の各第 2 位置のそれぞれのアミノ酸残基との間の推定距離を特徴付けており、距離マップクロップの生成は以下を含む：

シーケンス内の 1 つまたは複数の第 1 位置およびシーケンス内の 1 つまたは複数の第 2 位置を特定し、ここで、第 1 位置は、シーケンスの適切なサブセットであり、

シーケンス内の第 1 位置にあるアミノ酸残基とシーケンス内の第 2 位置にあるアミノ酸残基からネットワーク入力を決定し、

前記ネットワーク入力を距離予測ニューラルネットワークに提供し、該距離予測ニューラルネットワークは、距離予測ニューラルネットワーク重みの現在の値に従ってネットワーク入力を処理して、距離マップクロップを含むネットワーク出力を生成するよう構成され、

複数の距離マップクロップを使用して、所定のタンパク質の距離マップを生成する。

14. クレーム 1-13 の方法において、距離予測ニューラルネットワークは、シーケンスの第 1 位置にあるアミノ酸残基とシーケンスの第 2 位置にあるアミノ酸残基との間のねじれ角を特徴付ける補助出力を生成するようにトレーニングされている。

15. 所定のタンパク質の予測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、クレーム 1 -14 のいずれか一つに記載の方法において、

所定のタンパク質の予測構造は、複数の構造パラメータの値によって定義され、

予測構造を定義する複数の構造パラメータの初期値を取得し、

複数の更新反復のそれぞれにおいて、複数の構造パラメータの初期値を更新し、

距離マップを使用して、構造パラメータの現在の値によって定義された予測構造の品質を特徴付ける品質スコアを決定し、

複数の構造パラメータの 1 つまたは複数に関し、

構造パラメータの現在の値を調整して品質スコアを最適化し、

複数の更新反復の最後の更新反復の後、複数の構造パラメータの現在値によって定義される所定のタンパク質の予測構造を決定する。

16.コンピュータにより実装される方法において

（i）所定のタンパク質のアミノ酸シーケンス、および（ii）複数の構造パラメータの値によって定義される所定のタンパク質の予測構造を定義するデータを取得し、

所定のタンパク質のアミノ酸残基のシーケンスからネットワーク入力を決定し、

距離予測ニューラルネットワークの重みの現在の値に従って距離予測ニューラルネットワークを使用してネットワーク入力を処理し、所定のタンパク質の距離マップを生成し、

距離マップは、所定のタンパク質のアミノ酸残基のシーケンスにおける複数のアミノ酸残基のペアのそれぞれについて、所定のタンパク質の構造におけるアミノ酸残基のペア間の可能な距離範囲にわたるそれぞれの確率分布を定義しており、

距離マップによって定義された確率分布を使用して、所定のタンパク質の予測構造の品質を特徴付けるスコアを決定する。

33. リガンドを得る方法であって、リガンドは、薬物または工業用酵素のリガンドであり、該方法は以下を含む、

標的アミノ酸シーケンスを取得し、標的アミノ酸シーケンスは、標的タンパク質のアミノ酸シーケンスであり、

標的タンパク質の構造を決定するために、標的アミノ酸シーケンスを、アミノ酸またはアミノ酸残基のシーケンスとして使用して、クレーム 15 またはクレーム 32 に記載の方法を実行し、

1 つまたは複数の候補リガンドと標的タンパク質の構造との相互作用を評価し、

評価の結果に応じて、リガンドとして 1 つまたは複数の候補リガンドを選択する。

39. タンパク質のミスフォールディング疾患の存在を特定する方法であって、

タンパク質のアミノ酸シーケンスを取得し、

タンパク質の構造を決定するために、アミノ酸またはアミノ酸残基のシーケンスとしてタンパク質のアミノ酸配列を使用して、クレーム 15 またはクレーム 32 のいずれか

1 項に記載の方法を実行し、

ヒトまたは動物の体から得られたタンパク質のバージョンの構造を取得し、

タンパク質の構造を、人または動物の体から得られたバージョンのタンパク質の構造と比較し、

比較の結果に応じて、タンパク質のミスフォールディング疾患の存在を特定する。

4. AlphaFold に関する論文

AlphaFold に関する論文「ディープラーニングのポテンシャルを使用したタンパク質構造予測の改善」が Andrew W. Senior 氏らにより Nature 誌に発表されている[1]。

論文の図 1a には、AlphaFold の性能を示すグラフが示されている。

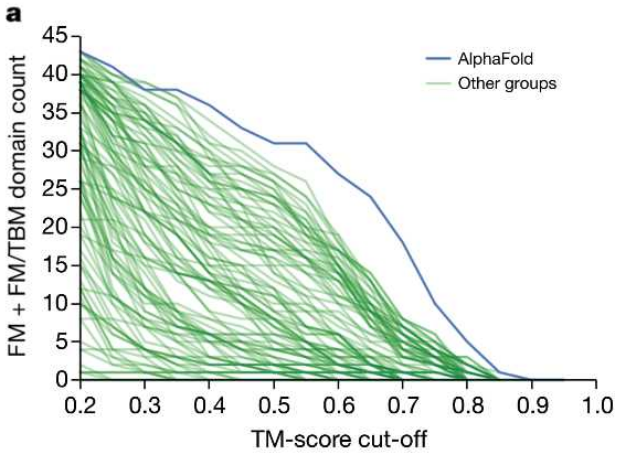


図 1 a

図 1a では、CASP13 コンテストに提出された他の構造予測システムのパフォーマンスと比較したものである。青色が AlphaFold であり、緑色が他のグループである。グラフの横軸は所定のテンプレートモデリング（TM : template modeling）スコアカットオフであり、縦軸は、予測されたフリーモデリング(FM: free modeling)タンパク質ドメインの数を示す。

FM タンパク質ドメインは、類似のタンパク質ドメインの構造が以前に（例えば、物理実験により）決定されていないドメインを指す。TM スコアは、タンパク質の本来の（すなわち、実際の）構造に対するタンパク質の提案された構造の骨格形状の一致の程

[1] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu & Demis Hassabis“Improved protein structure prediction using potentials from deep learning”2020 年 1 月 15 日 Nature 誌

度を測定する 0 と 1 の間のスコアを指す。図 1a に示すように、AlphaFold は、ほぼすべての TM スコアカットオフで他の構造予測システムより優れている。

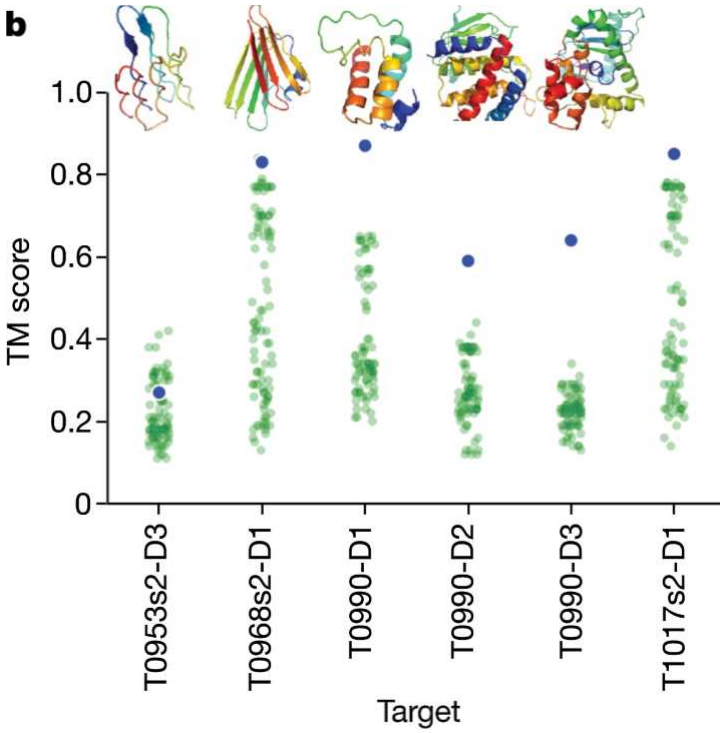


図 1b

図 1b は、新たに決定された 6 つのタンパク質構造(横軸)について、AlphaFold によって生成された構造予測の TM スコア（青丸)、および、CASP13 に提出された他の構造予測システムによって生成された構造予測の TM スコア（緑丸）を示す。図 1b を参照して説明した構造予測システムは、一般的に他の構造予測システムより優れている。

a

Contact precisions		L long			L/2 long			L/5 long		
Set	N	AF	498	032	AF	498	032	AF	498	032
FM	31	**46.1**	43.1	40.1	**58.5**	54.9	51.6	**69.9**	67.3	61.9
FM/TBM	12	**59.1**	53.0	48.9	**74.2**	64.5	64.2	**85.3**	81.0	79.6
TBM	61	**68.3**	65.5	61.9	**82.4**	80.3	76.4	**90.6**	90.5	87.1

図 2a

図 2a は、AlphaFold を使用することによって達成することができる性能利得の例を示す。図 2a では、最も可能性の高い L、L/2、または L/5 のアミノ酸残基の接触に対する CASP13 での長距離接触予測の精度を示している。ここで、L はドメインの長さ

である。

距離予測システムによって生成されるアミノ酸のペア間の距離範囲にわたる確率分布(AF)は、接触予測に対して閾値処理され、CASP13 の 2 つの最高ランクの接触予測方法 032（TripletRes）および 498（RaptorX-Contact）と比較される。

図 2a では、フリーモデリング(FM) タンパク質ドメイン、テンプレートベースモデリング(TBM) タンパク質ドメイン（類似したシーケンスを持つタンパク質ドメインが既知構造を有している場合)、及び、中間の FM / TBM タンパク質ドメインに関し、距離予測システムの接触予測精度を示している。 本距離予測システムは、一般的に他の接触予測システムよりも優れていることが理解できる。

次に AlphaFold のネットワーク構成について説明する。

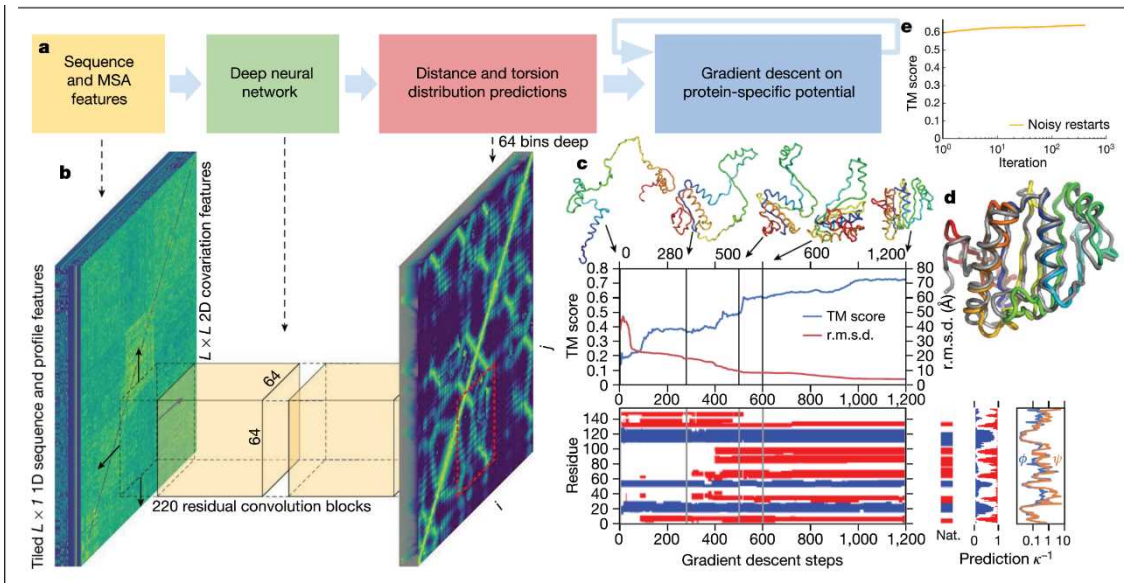


図 3

図 3 は構造予測システムを使用して、タンパク質の予測構造を決定するためのデータフローを示している。L x L 2D 共分散特徴とタイル型 L x 1 1D シーケンスおよびプロファイル特徴（L はアミノ酸シーケンスの長さ）とを連結して、シーケンスと MSA(multiple sequence alignment)特徴（図 3a）を生成する。

シーケンスと MSA 特徴は、数値の 3D 配列として表すことができる。当該特徴(図 3a)からの 64×64 クロップは、220 の残差畳み込みブロックを有する距離予測ニューラルネットワークを使用して処理され、完全距離マップのクロップを生成する。

完全距離マップのクロップは融合され（例えば、平均化され）、完全距離マップを生成する。完全距離マップは、タンパク質内のアミノ酸の各ペア間の 64 の可能な距離範囲にわたるそれぞれの確率分布を特定する。

距離予測ニューラルネットワークの個別の出力ヘッドは、タンパク質の各アミノ酸の構造パラメータ分布（たとえば、ねじれ角分布）を生成する（つまり、この例では、構造予測ネットワークと距離予測ネットワークはいくつかのパラメータ値を共有する）。タンパク質の予測構造を定義する構造パラメータの初期値は、勾配降下法の複数の反復にわたって更新され、タンパク質の最終予測構造が生成される。

各反復で、構造パラメータの現在の値によって定義された予測構造の品質スコアは、（i）距離マップ、（ii）構造パラメータ分布、および（iii）予測される構造に関連する原子間ポテンシャルエネルギーを特徴付けるファンデルワールスポテンシャル(van der Waals potential)に基づいた物理スコア、に基づいて決定される。

品質スコアは、構造パラメータの現在の値に関して微分可能であり、品質スコアの勾配は、構造パラメータの現在の値に関して決定される。勾配降下最適化手法は、構造パラメータの更新された値を決定するために、品質スコアの勾配を使用して構造パラメータの現在の値を調整するために使用される。

グラフ図 3c は、各勾配降下ステップにおけるタンパク質の予測構造と実際の構造との間の TM スコア(青色)および RMSD(赤色（Root Mean Square Deviation 平均二乗誤差)を示す。タンパク質の予測構造が、勾配降下ステップのシーケンスにわたって、タンパク質の実際の構造をより正確に近似していることが理解できる。

(i)最後の勾配降下ステップ後のタンパク質の予測構造、および(ii)タンパク質の実際の構造のオーバーレイの 3D 視覚化(図 3c 及び d)から、タンパク質の最終予測構造がタンパク質の実際の構造に正確に近似していることが理解できる。

グラフ図 3e は、さまざまな初期化で勾配降下ステップを複数回実行してさまざまな予測構造を生成し、最適な予測構造をタンパク質の最終予測構造として選択することで達成できる TM スコアの改善を示している。

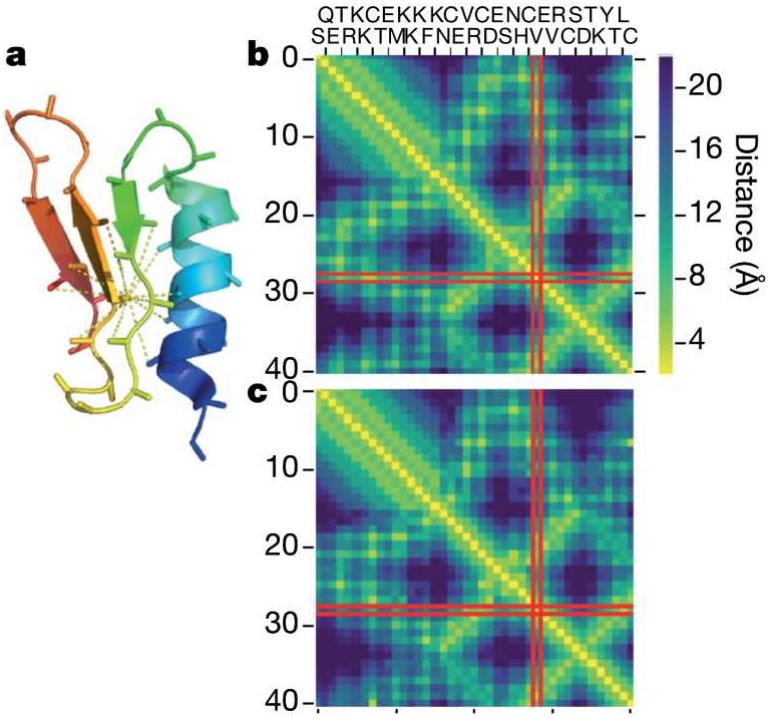


図 4

図 4 は、距離予測システムによって生成されたタンパク質の距離マップの態様を示す。図 4b は、タンパク質の実際の（すなわち、ネイティブな）残基間距離を示す距離マップである。図 4c は、距離予測システムを使用して生成され、タンパク質の残基間距離範囲にわたる確率分布のモードを示す距離マップである。図 4c に示す予測距離マップが、図 4b に示す実際の距離マップ 1902 の正確な近似であることが理解できる。

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「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 2.0」がある。