AI 特許紹介(68)
AI 特許を学ぶ！究める！
～AlphaMissense 特許～

2024 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　2023 年 10 月 11 日
公開日　2024 年 4 月 18 日
公開番号 WO2024079204
発明の名称　アミノ酸スコア分布を用いたタンパク質変異の病原性予測

204 特許は、タンパク質の多重配列アライメント（MSA：Multiple Sequence Alignment）を表す MSA 表現を病原性予測ニューラルネットワークに入力することにより、病原性スコアを生成する AlphaMissense 技術に関する。

2.特許内容の説明
図１は、病原性予測システム 120 である。

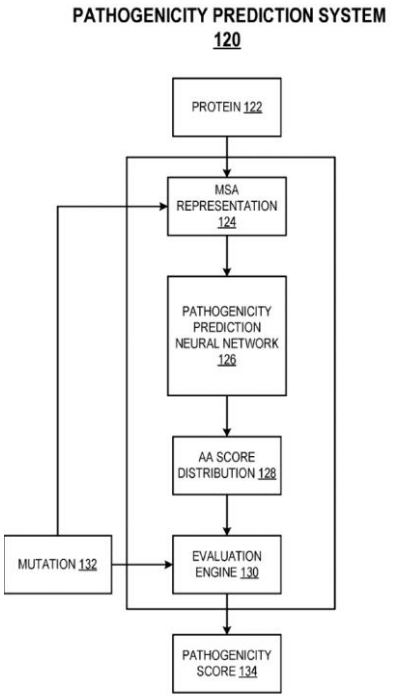


FIG. 1

病原性予測システム 120 は、(i)タンパク質 122 のアミノ酸配列、および (ii)タンパク質 122 のアミノ酸配列に対する変異 132 を定義するデータを処理し、病原性スコア 134 を生成する。病原性スコア 134 は、突然変異 132 が病原性突然変異である可能性を示す。変異 132 は、(i)タンパク質 122 のアミノ酸配列における変異位置、および(ii)アミノ酸配列における変異位置の代替アミノ酸を指定する。変異 132 は、例えば、タンパク質 122 をコードする遺伝子におけるミスセンス変異の結果である。病原性スコア 134 は、病原性スコアの値が高いほど、タンパク質変異 122 が病原性である可能性が高いことを表す。

病原性予測システム 120 は、病原性予測ニューラルネットワーク 126 と評価エンジン 130 を含む。変異 132 の病原性スコア 134 を生成するために、病原性予測システム 120 は、タンパク質 122 の多重配列アライメント(MSA)を定義するデータを取得し、タンパク質 122 の MSA を表す MSA 表現 124 を生成する。MSA 表現 124 は、MSA に含まれる各タンパク質のアミノ酸配列の各位置に対応するそれぞれの埋め込みを含む埋め込みのコレクションを含む。

また、病原性予測システム 120 は、（参照）タンパク質のアミノ酸配列内の変異位置（すなわち、変異 132 によって指定される）に対応する MSA 表現 124 内の埋め込みを「マスク」することができる。

タンパク質の MSA は、タンパク質の変異の病原性を予測するのに関係する情報をエンコードする。たとえば、タンパク質のアミノ酸配列の各位置について、MSA は、タンパク質の進化の歴史の過程でその位置のアミノ酸が異なるアミノ酸に変化した頻度を示す情報をエンコードする。タンパク質のアミノ酸配列のある位置のアミノ酸が、タンパク質の進化の歴史の中で頻繁に変異した場合、その位置の変異が病原性である可能性は低くなる。逆に、タンパク質のアミノ酸配列のある位置のアミノ酸が、タンパク質の進化の歴史の中でめったに変異しなかった場合、その位置の変異が病原性である可能性は高くなる。直感的に、タンパク質のアミノ酸配列のある位置の変異が病原性を引き起こす場合、進化圧力によって、その位置の変異を含むアミノ酸配列がタンパク質の進化の歴史の中であまり出現しなくなる可能性がある。

病原性予測システム 120 は、病原性予測ニューラルネットワーク 126 を使用して MSA 表現 124 を含むネットワーク入力を処理し、アミノ酸のセットのスコア分布 128 を生成する。評価エンジン 130 は、病原性予測ニューラルネットワーク 126 によって生成されたアミノ酸スコア分布 128 に基づいて、変異位置の変異 132 の病原性スコア 134 を生成する。たとえば、評価エンジン 130 は、(i)スコア分布 128 に基づく（つまり、スコア分布 128 によって与えられた）変異位置の元のアミノ酸のスコアと、(ii)スコア分布 128 に基づく、変異位置の置換アミノ酸のスコアとの差の尺度を決定するなどして、変異 132 の病原性スコア 134 を生成する。

変異位置の元のアミノ酸とは、タンパク質 122 の元の（つまり、変異していない）アミノ酸配列内の変異位置にあるアミノ酸を指す。つまり、評価エンジン 130 は、前述のように、変異位置にマスクされた埋め込みを持つ MSA 表現を処理してスコア分布 128 を決定し、スコア分布内の元のアミノ酸と置換アミノ酸のスコアから、たとえばこれらのスコアの差から、変異 132 の病原性スコア 134 を生成する。

評価エンジン 130 は、病原性スコア 134 を使用して、突然変異 132 を「病原性」または「良性」に分類する。たとえば、評価エンジン 130 は、突然変異 132 の病原性スコア 134 がしきい値を満たす（たとえば、しきい値を超える）場合、突然変異 132 を病原性として分類する。評価エンジン 130 は、突然変異の病原性スコア 134 がしきい値を満たさない（たとえば、しきい値を超えない）場合、突然変異 132 を良性として分類する。

図 3 は、トレーニングシステム 142 の一例を示す。

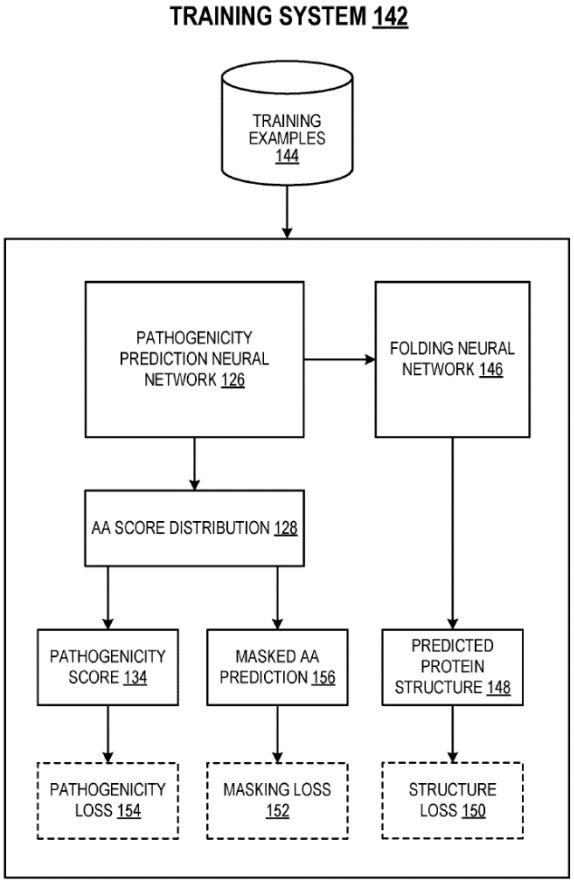


FIG. 3

トレーニングシステム 142 は、一連のトレーニンサンプル 144 で病原性予測ニューラルネットワーク 126 をトレーニングする。病原性予測ニューラルネットワークは、病原性予測タスクを実行するようにトレーニングされるが、タンパク質のマスク解除タスク及び、タンパク質構造予測タスクをも実行するようにトレーニングされる。

トレーニングシステム 142 は、病原性予測ニューラルネットワーク 126 を、(i) 病原性損失 154 を最適化するための病原性トレーニングサンプル、(ii) マスキング損失 152 を最適化するためのマスクトレーニングサンプル、および (iii) 構造損失 150 を最適化するための構造トレーニングサンプルに基づいてトレーニングする。

病原性トレーニング例に基づいて病原性予測ニューラルネットワーク 126 をトレーニングするために、トレーニングシステム 142 は、対応するトレーニングタンパク質を特徴付けるネットワーク入力を病原性予測ニューラルネットワーク 126 に提供する。トレーニングサンプルによって指定された各変異について、トレーニングシステム 142 は、トレーニングタンパク質のアミノ酸配列内の変異の変異位置に対応する埋め込みを

マスクする。

トレーニングシステム 142 は、病原性予測ニューラルネットワーク 126 を使用してトレーニングタンパク質のネットワーク入力を処理し、トレーニングサンプルで指定された各変異の変異位置のそれぞれのアミノ酸スコア分布を生成する。トレーニングサンプルで指定された各変異について、トレーニングシステムは、病原性予測ニューラルネットワーク 126 によって変異位置に対して生成されたアミノ酸スコア分布に基づいて、それぞれの予測病原性スコア 134 を生成する。

トレーニングシステム 142 は、病原性トレーニングサンプルによって指定された各突然変異の病原性損失 154 を評価し、次に、各突然変異の病原性損失の組み合わせ（たとえば、平均または合計）として、病原性トレーニングサンプルの全体的な病原性損失を決定する。トレーニングシステム 142 は、(i) 突然変異の予測病原性スコアと (ii) 突然変異のターゲット病原性スコアとの間の誤差（たとえば、ヒンジ損失、クロスエントロピー損失、またはその他の適切な損失）に基づいて突然変異の病原性損失を評価する。

トレーニング システム 142 は、病原性予測ニューラルネットワーク 126 のパラメータに対する病原性損失 154 の勾配を、例えばバックプロパゲーションを使用して決定する。次に、トレーニングシステム 142 は、勾配を使用して、例えば適切な勾配降下法最適化技術を使用して、病原性予測ニューラルネットワーク 126 のパラメータの現在の値を調整する。

病原性予測ニューラル ネットワーク 126 をマスクされたトレーニングサンプルでトレーニングするために、トレーニングシステムは、病原性予測ニューラルネットワーク 126 に対応するトレーニングタンパク質を特徴付けるネットワーク入力を提供する。トレーニングシステム 142 は、マスクされたトレーニング例によってマスクされた位置として指定されたトレーニングタンパク質のアミノ酸配列内の位置に対応する MSA 表現 124 内の各埋め込みをマスクする。

次に、トレーニングシステム 142 は、病原性予測ニューラルネットワーク 126 を使用してトレーニングタンパク質のネットワーク入力を処理し、トレーニングタンパク質のアミノ酸配列内の各マスク位置に対応するそれぞれのアミノ酸スコア分布を生成する。トレーニングシステム 142 は、各マスク位置のマスキング損失を評価し、次に、各マスク位置のマスキング損失の組み合わせ（たとえば、平均または合計）としてトレーニングサンプルの全体的なマスキング損失を決定する。トレーニングシステム 142 は、トレーニングタンパク質内のマスク位置のアミノ酸の、マスク位置のスコア分布の下で

のスコアに基づいて、たとえば、マスキングによるスコアの変化に基づいて、マスク位置のマスキング損失を評価する。マスキング損失は、各マスク位置について病原性予測ニューラルネットワークによって生成されたそれぞれの予測の精度を測定できる。

トレーニングシステム 142 は、例えばバックプロパゲーションを使用して、病原性予測ニューラルネットワーク 126 のパラメータに対するマスキング損失 152 の勾配を決定する。次に、トレーニングシステム 142 は、勾配を使用して、例えば適切な勾配降下法最適化手法を使用して、病原性予測ニューラルネットワーク 126 のパラメータの現在の値を調整する。

病原性予測ニューラルネットワーク 126 をマスキング損失 152 でトレーニングすることにより、トレーニングシステム 142 は、病原性予測ニューラルネットワーク 126 がアンマスキングタスクを実行するようにトレーニングする。アンマスキングタスクでは、病原性予測ニューラルネットワークが、アミノ酸配列の残りのマスクされていない部分からのコンテキスト情報に基づいて、タンパク質のアミノ酸配列内のマスクされた位置にあるアミノ酸の ID をデコードする必要がある。アンマスキングタスクを効果的に実行することを学習すると、病原性予測ニューラルネットワークのパラメータ値にタンパク質生化学の理解を暗黙的にエンコードすることができ、それによって病原性予測のタスクにおける病原性予測ニューラルネットワーク 126 のパフォーマンスを向上させることができる。

病原性予測ニューラルネットワーク 126 を構造トレーニングサンプルでトレーニングするために、トレーニングシステム 142 は、対応するトレーニングタンパク質を特徴付けるネットワーク入力を病原性予測ニューラルネットワーク 126 に提供する。トレーニングシステム 142 は、病原性予測ニューラルネットワーク 126 を使用してトレーニングタンパク質のネットワーク入力を処理し、病原性予測ニューラルネットワークの中間出力をフォールディングニューラルネットワーク 146 に提供する。フォールディングニューラルネットワーク 146 は、病原性予測ニューラルネットワーク 126 の中間出力を処理して、トレーニングタンパク質の予測構造 148 を定義するデータを生成する。

フォールディングニューラルネットワーク 146 は、病原性予測ニューラルネットワーク 126 の任意の適切な中間出力を受信する。たとえば、病原性予測ニューラルネットワーク 126 には、埋め込みニューラルネットワークを含めることができ、フォールディングニューラルネットワーク 146 は、埋め込みニューラルネットワークの出力を受信する。フォールディングニューラルネットワークは、埋め込みニューラルネットワーク

によって生成された更新された MSA 表現または更新されたペア埋め込みを受信する。

構造トレーニング例の構造損失 150 は、(i)フォールディングニューラルネットワーク 146 によって生成された予測タンパク質構造 148 と、(ii)構造トレーニングサンプルによって指定されたターゲットタンパク質構造との間の誤差を測定する。トレーニングシステム 142 は、例えばバックプロパゲーションを使用して、病原性予測ニューラルネットワーク 126 およびフォールディングニューラルネットワーク 146 のパラメータに対する構造損失 150 の勾配を決定する。次に、トレーニングシステム 142 は、例えば適切な勾配降下最適化技術によって、勾配を使用して病原性予測ニューラルネットワーク 126 およびフォールディングニューラルネットワーク 146 のパラメータの現在の値を調整する。つまり、トレーニングシステム 142 は、構造損失 150 の勾配を、折り畳みニューラルネットワーク 146 を介して病原性予測ニューラルネットワーク 126 に逆伝播する。

病原性予測ニューラルネットワーク 126 のトレーニング後には、折り畳みニューラルネットワーク 146 は不要である。したがって、病原性予測ニューラルネットワーク 126 のトレーニング後には、折り畳みニューラルネットワーク 146 を破棄することができる。

病原性予測ニューラルネットワーク 126 およびフォールディングニューラルネットワーク 146 を構造損失 150 でトレーニングすることにより、トレーニングシステム 142 は、病原性予測ニューラルネットワーク 126 およびフォールディングニューラルネットワーク 146 をトレーニングして、タンパク質構造予測のタスクを共同で実行できるようにする。タンパク質構造予測タスクを効果的に実行するための学習は、病原性予測ニューラルネットワークのパラメータ値にタンパク質生化学の理解を暗黙的にエンコードすることができ、それによって病原性予測タスクにおける病原性予測ニューラルネットワークのパフォーマンスを向上させることができる。

3.クレーム

204 特許のクレーム 1 は以下の通りである。

1. 1 台以上のコンピュータによって実行される方法において、

タンパク質の変異が病原性変異である可能性を特徴付ける病原性スコアを生成し、

変異は、タンパク質のアミノ酸配列中の変異位置において元のアミノ酸を置換アミノ酸に置き換えることによってタンパク質のアミノ酸配列を改変するものであり、

病原性スコアの生成は以下を含む:

病原性予測ニューラルネットワークへのタンパク質の多重配列アライメント(MSA) を表す MSA 表現を含むネットワーク入力を生成し、

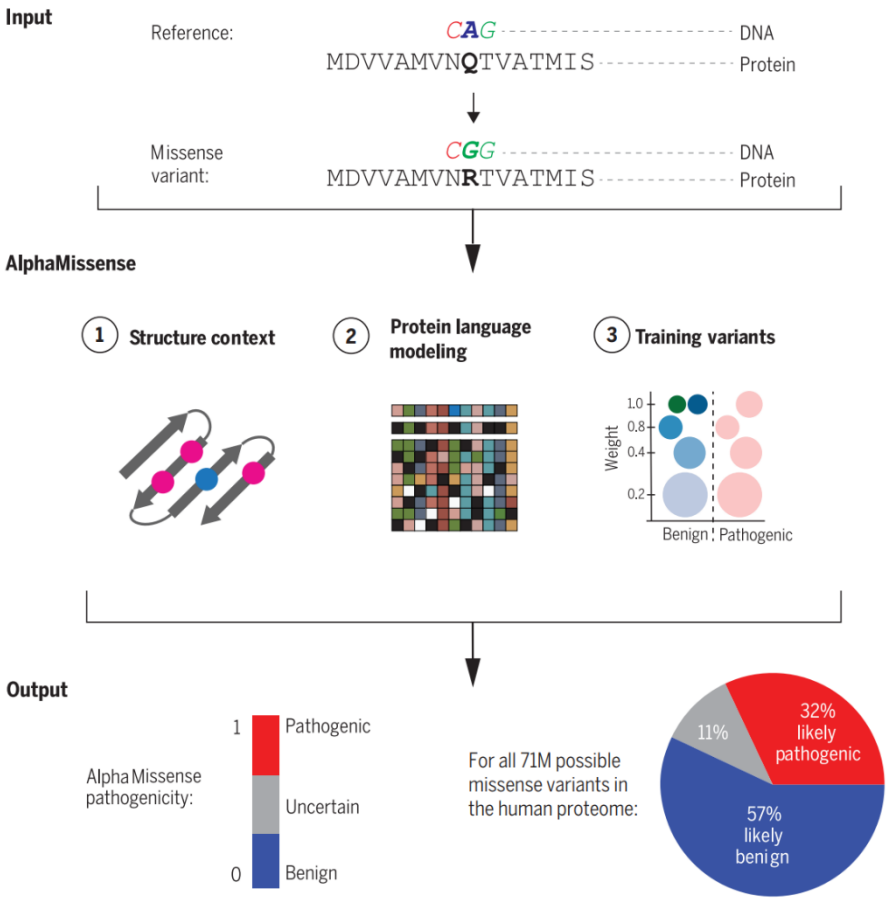
アミノ酸セットのスコア分布を生成するために、病原性予測ニューラルネットワークを使用してネットワーク入力を処理し、

(i)スコア分布における元のアミノ酸のスコアと、(ii)スコア分布における置換アミノ酸のスコアとの差に基づいて病原性スコアを生成する。

4. 本特許に関連する論文

本特許に関する論文“Accurate proteome-wide missense variant effect prediction with AlphaMissense”[1]が、Deepmind の Jun Cheng 氏らにより公表されている。

下記図は AlphaMissense による予測プロセスを示す説明図である。

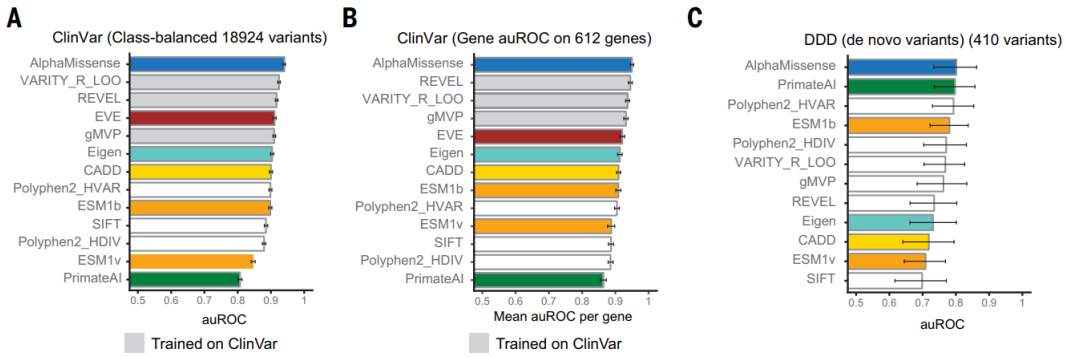


AlphaMissense はミスセンス変異を入力として受け取り、その病原性を予測する。

[1] Jun Cheng, et al. “Accurate proteome-wide missense variant effect prediction with AlphaMissense” Science 381, 1303 (2023) 22 September 2023

ヒトと霊長類の変異体集団頻度データに基づいて AlphaFold をファインチューニングすることにより、既知の疾患変異体の信頼性が較正されている。AlphaMissense は、ミスセンス変異体が病原性である可能性を予測し、それを良性である可能性が高い、病原性である可能性が高い、または不確実のいずれかに分類する。

下記グラフは、臨床的にキュレートされた分類ベンチマークにおける AlphaMissense のパフォーマンスを示す。



ベンチマークは、受信者操作曲線の下の面積（auROC: area under the receiver operator curve）によって評価される。いずれのケースにおいても AlphaMissense が良い結果を示している。

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、生成 AI ビジネスコース修了。

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 3」、「ブロックチェーン 3.0(共著)」がある。