

(19)日本国特許庁(JP)

(12)公表特許公報(A)

(11)公表番号

特表2023-534616

(P2023-534616A)

(43)公表日 令和5年8月10日(2023.8.10)

(51)国際特許分類	F I	テーマコード (参考)
G 1 6 H 50/00 (2018.01)	G 1 6 H 50/00	5 L 0 9 9
G 1 6 B 20/20 (2019.01)	G 1 6 B 20/20	

審査請求 未請求 予備審査請求 未請求 (全58頁)

(21)出願番号	特願2022-578715(P2022-578715)	(71)出願人	500358711 イルミナ インコーポレイテッド アメリカ合衆国 カリフォルニア州 9 2 1 2 2 サンディエゴ イルミナ ウェイ 5 2 0 0
(86)(22)出願日	令和3年7月21日(2021.7.21)	(74)代理人	100108453 弁理士 村山 靖彦
(85)翻訳文提出日	令和4年12月19日(2022.12.19)	(74)代理人	100110364 弁理士 実広 信哉
(86)国際出願番号	PCT/US2021/042599	(74)代理人	100133400 弁理士 阿部 達彦
(87)国際公開番号	WO2022/020487	(72)発明者	ホン・ガオ アメリカ合衆国・カリフォルニア・9 2 1 2 2・サン・ディエゴ・イルミナ・ウ エイ・5 2 0 0
(87)国際公開日	令和4年1月27日(2022.1.27)		
(31)優先権主張番号	63/055,724		
(32)優先日	令和2年7月23日(2020.7.23)		
(33)優先権主張国・地域又は機関	米国(US)		
(81)指定国・地域	AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA ,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,A T,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR ,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC, 最終頁に続く		最終頁に続く

(54)【発明の名称】 変異体病原性スコアリング及び分類、並びにそれらの使用

(57)【要約】

遺伝子変異体の病原性スコア(206)の導出及び使用が本明細書に記載されている。病原性スコアリングプロセスの応用、使用、及び変形には、変異体を病原性又は良性として特徴付けるための閾値(212、218)の導出及び使用、遺伝子変異体に関連付けられた選択効果の推定、病原性スコア(206)を使用した遺伝的疾患の有病率の推定、並びに病原性スコア(206)を評価するために使用される方法の再較正が含まれるが、これらに限定されない。

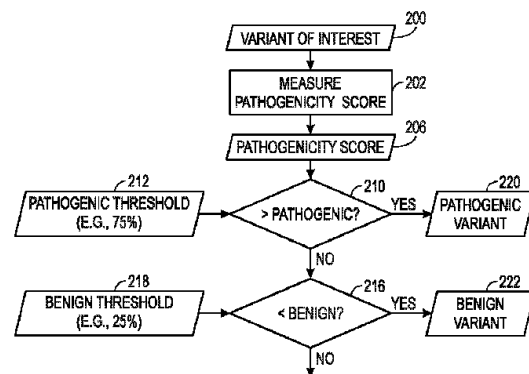


FIG. 7

【特許請求の範囲】

【請求項 1】

関心のある変異体（200）を分類するための方法であって、前記方法は、

コンピュータシステムの1つ以上のプロセッサ上で、前記関心のある変異体（200）を分類するための保存された命令を実行することを含み、前記保存された命令が、実行されると、前記1つ以上のプロセッサに、

遺伝子の関心のある変異体（200）の病原性スコア（206）を入力として受信することと、

前記関心のある変異体（200）の前記病原性スコア（206）を遺伝子固有病原性閾値（212）と比較することと、

前記病原性スコア（206）が前記遺伝子固有病原性閾値（212）を超えていることに応答して、前記関心のある変異体（200）を病原性であるとして分類することと（210）、を含む行為を実行させる、方法。

【請求項 2】

前記病原性スコア（206）が、前記関心のある変異体（200）についてのヒト及びヒト以外の霊長類における浄化選択の程度に基づく、請求項1に記載の方法。

【請求項 3】

病原性変異体の平均病原性スコアと、病原性スコアのパーセンタイルとの間の相関を決定すること、を更に含み、

病原性変異体の前記平均病原性スコアと、病原性スコアの前記パーセンタイルとの間の前記相関が、前記遺伝子固有病原性閾値（212）を定義するために使用される、請求項1に記載の方法。

【請求項 4】

前記病原性スコア（206）が、アミノ酸配列から前記病原性スコア（206）を生成するように訓練されたニューラルネットワーク（102）によって処理されたアミノ酸配列に基づいて生成される、請求項1に記載の方法。

【請求項 5】

前記アミノ酸配列の中央アミノ酸が、前記関心のある変異体（200）に対応する、請求項4に記載の方法。

【請求項 6】

前記ニューラルネットワーク（102）が、ヒト配列及びヒト以外の配列の両方を使用して訓練されている、請求項4に記載の方法。

【請求項 7】

前記遺伝子固有病原性閾値（212）が、51パーセンタイル以上かつ99パーセンタイル以下によって定義される範囲内にある、請求項1に記載の方法。

【請求項 8】

前記遺伝子固有病原性閾値（212）が、75パーセンタイル以上かつ99パーセンタイル以下によって定義される範囲内に含まれる、請求項1に記載の方法。

【請求項 9】

前記保存された命令が、実行されると、前記1つ以上のプロセッサに、

前記関心のある変異体（200）の前記病原性スコア（206）を遺伝子固有良性閾値（218）と比較することと、

前記病原性スコア（206）が前記遺伝子固有良性閾値（218）未満であることに応答して、前記関心のある変異体（200）を良性であるとして分類することと、を含む更なる行為を実行させる、請求項1に記載の方法。

【請求項 10】

良性変異体の平均病原性スコアと、病原性スコアのパーセンタイルとの間の相関を決定すること、を更に含み、

良性変異体の前記平均病原性スコアと、病原性スコアの前記パーセンタイルとの間の前記相関が、前記遺伝子固有良性閾値（218）を定義するために使用される、請求項9に

10

20

30

40

50

記載の方法。

【請求項 1 1】

前記遺伝子固有良性閾値（218）が、1パーセンタイル以上かつ49パーセンタイル以下によって定義される範囲内にある、請求項9に記載の方法。

【請求項 1 2】

前記遺伝子固有良性閾値（218）が、1パーセンタイル以上かつ25パーセンタイル以下によって定義される範囲内にある、請求項9に記載の方法。

【請求項 1 3】

プロセッサ実行可能命令を保存する非一時的コンピュータ可読媒体であって、前記プロセッサ実行可能命令が、1つ以上のプロセッサによって実行されると、前記1つ以上のプロセッサに、

病原性スコアリングニューラルネットワーク（102）を使用して遺伝子の関心のある変異体（200）を処理して、前記関心のある変異体（200）の病原性スコア（206）を生成するステップと、

前記関心のある変異体（200）の前記病原性スコア（206）を、前記遺伝子に固有の1つ以上の閾値と比較するステップと、

前記病原性スコア（206）と、前記遺伝子に固有の前記1つ以上の閾値との比較に基づいて、前記関心のある変異体（206）を、病原性であるか、良性であるか、又は病原性でも良性でもないとして分類するステップと、を含むステップを実行させる、非一時的コンピュータ可読媒体。

【請求項 1 4】

前記1つ以上の閾値が遺伝子固有病原性閾値（212）を含み、前記病原性スコア（206）が前記遺伝子固有病原性閾値（212）を超えていることに応答して、前記関心のある変異体（200）が病原性であるとして分類される、請求項13に記載の非一時的コンピュータ可読媒体。

【請求項 1 5】

前記1つ以上の閾値が遺伝子固有良性閾値（218）を含み、前記病原性スコア（206）が前記遺伝子固有良性閾値（218）を下回っていることに応答して、前記関心のある変異体（200）が良性であるとして分類される、請求項13に記載の非一時的コンピュータ可読媒体。

【請求項 1 6】

前記遺伝子に固有の前記1つ以上の閾値が、遺伝子固有病原性閾値（212）及び遺伝子固有良性閾値（218）を含み、前記病原性スコア（206）が前記遺伝子固有良性閾値（218）と前記遺伝子固有病原性閾値（212）との間にあることに応答して、前記関心のある変異体（200）が病原性でも良性でもないとして分類される、請求項13に記載の非一時的コンピュータ可読媒体。

【請求項 1 7】

前記関心のある変異体（200）の前記病原性スコア（206）を前記遺伝子に固有の1つ以上の閾値と比較することが、前記関心のある変異体（200）の前記病原性スコア（206）を、第2の閾値の前に第1の閾値と比較することを含み、前記第2の閾値との比較は、前記第1の閾値との前記比較の結果次第である、請求項13に記載の非一時的コンピュータ可読媒体。

【請求項 1 8】

前記関心のある変異体（200）の前記病原性スコア（206）を前記遺伝子に固有の1つ以上の閾値と比較することが、前記関心のある変異体（200）の前記病原性スコア（206）を、第1の閾値及び第2の閾値と並列して比較することを含む、請求項13に記載の非一時的コンピュータ可読媒体。

【請求項 1 9】

変異体分類システムであって、

プロセッサ実行可能命令を符号化する1つ以上のメモリ構造と、

10

20

30

40

50

前記 1 つ以上のメモリ構造と通信する 1 つ以上のプロセッサと、を備え、前記 1 つ以上のプロセッサは前記命令を実行して、

遺伝子の関心のある変異体（200）の病原性スコア（206）を入力として受信することと、

前記関心のある変異体（200）の前記病原性スコア（206）を、前記遺伝子に固有の 1 つ以上の閾値と比較することと、

前記病原性スコア（206）と、前記遺伝子に固有の前記 1 つ以上の閾値との比較に基づいて、前記関心のある変異体（200）を、病原性であるか、良性であるか、又は病原性でも良性でもないとして分類することと、を含む動作を実行させる、変異体分類システム。

10

【請求項 20】

前記 1 つ以上の閾値が遺伝子固有病原性閾値（212）を含み、前記病原性スコア（206）が前記遺伝子固有病原性閾値（212）を超えていることに応答して、前記関心のある変異体（200）が病原性であるとして分類される、請求項 19 に記載の変異体分類システム。

【請求項 21】

前記 1 つ以上の閾値が遺伝子固有良性閾値（218）を含み、前記病原性スコア（206）が前記遺伝子固有良性閾値（218）を下回っていることに応答して、前記関心のある変異体（200）が良性として分類される、請求項 19 に記載の変異体分類システム。

【発明の詳細な説明】

20

【技術分野】

【0001】

（関連出願の相互参照）

本出願は、2020 年 7 月 23 日に提出された米国仮特許出願第 63 / 055, 724 号の優先権を主張し、この仮出願は、参照によりその全体が本明細書に組み込まれる。

【0002】

（発明の分野）

開示される技術は、生物学的配列変異体の病原性を評価し、病原性評価を使用して他の病原性関連データを導出するためにコンピュータ及びデジタルデータ処理システムに実装され得る、人工知能と称されることがある機械学習技術の使用に関する。これらの手法は、知能をエミュレーションするための対応するデータ処理方法及び製品（すなわち、知識ベースのシステム、推測システム及び知識取得システム）、並びに／又は、不確実性を用いて推測するためのシステム（例えば、ファジー論理システム）、適応システム、機械学習システム、及び人工ニューラルネットワークを含むか、又は利用する。特に、開示される技術は、病原性評価、及びそのような病原性情報の使用又は改良のための深層学習ニューラルネットワークを訓練するために深層学習ベースの技術を使用することに関する。

30

【背景技術】

【0003】

このセクションで考察される主題は、単にこのセクションにおける言及の結果として、先行技術であると想定されるべきではない。同様に、このセクションで言及した問題、又は背景として提供された主題と関連付けられた問題は、先行技術において以前に認識されていると想定されるべきではない。このセクションの主題は、単に、異なるアプローチを表し、それ自体はまた、特許請求される技術の実施態様に対応し得る。

40

【0004】

遺伝的変異は多くの疾患を説明するのに役立ち得る。全てのヒトは固有の遺伝暗号を有し、個体の群内には多くの遺伝的変異体が存在する。病的な遺伝的変異体の多く又はほとんどは、自然選択によってゲノムから消失してきた。しかし、どの遺伝的変異が臨床的に重要である可能性が高いかを特定することは依然として困難である。

【0005】

50

更に、変異体の特性及び機能的作用（例えば、病原性）をモデル化することは、ゲノムの分野において困難なタスクである。機能的ゲノム配列決定技術の急速な進歩にもかかわらず、細胞型特異的転写調節システムの複雑さのために、変異体の機能的帰結の解釈は大きな課題である。

【発明の概要】

【0006】

変異体病原性分類器を構築し、そのような病原性分類器の情報を使用又は改良するためのシステム、方法、及び製造品が記載される。そのような実装形態は、本明細書に記載のシステム及び方法論のアクションを実行するためにプロセッサによって実行可能な命令が保存されている非一時的コンピュータ可読記憶媒体を含むか、又は利用することができる。明示的に列挙又は記載されていない場合であっても、実装形態の1つ以上の特徴が、基本実装形態又は他の実装形態と組み合わせられ得る。更に、相互排他的ではない実装形態は組み合わせ可能であると教示され、実装形態の1つ以上の特徴を他の実装形態と組み合わせることができる。本開示では、これらのオプションについてユーザは定期的に喚起され得る。しかし、これらのオプションを繰り返す記載のいくつかの実装形態からの省略は、以下のセクションで教示される潜在的な組み合わせを限定するものとして解釈されるべきではない。代わりに、これらの記載は、参照により以下の各実装形態に組み込まれる。

10

【0007】

開示されるこのシステムの実装形態及び他のシステムは、任意選択で、本明細書で議論される特徴の一部又は全てを含む。システムはまた、開示される方法に関連して説明される特徴を含み得る。簡潔さのために、システム特徴の代替の組み合わせは、個別に列挙されてはいない。また、システム、方法、及び製造物品に適用可能な特徴は、基本特徴の各法定分類セットに対して繰り返されない。読者は、特定された特徴が他の法定分類における基本特徴とどのように容易に組み合わせることができるかを理解するであろう。

20

【0008】

議論される主題の一態様では、メモリに結合された多数のプロセッサ上で実行される、畳み込みニューラルネットワークベースの変異体病原性分類器を訓練する方法論及びシステムが記載される。あるいは、他のシステム実装形態では、ニューラルネットワークベースの分類器に加えて、又はその代わりに、訓練又は好適にパラメータ化された統計モデル若しくは技術及び／又は他の機械学習手法を用いることができる。システムは、良性の変異体及び病原性変異体から生成されたタンパク質配列ペアの良性訓練例及び病原性訓練例を使用する。良性変異体は、コモンヒトミスセンス変異体、及び、一致する参照コドン配列をヒトと共有する代替の非ヒト霊長類のコドン配列で発生する非ヒト霊長類ミスセンス変異体を含む。サンプリング対象のヒトは、アフリカ人／アフリカ系アメリカ人（AFRと略される）、アメリカ人（AMRと略される）、アシュケナージ系ユダヤ人（ASJと略される）、東アジア人（EASと略される）、フィンランド人（FINと略される）、非フィンランド系ヨーロッパ人（NFEと略される）、南アジア人（SASと略される）、及びその他（OTHと略される）を含み得るか、又はそのように特徴付けられ得る異なるヒト亜集団に属し得る。非ヒト霊長類ミスセンス変異体は、限定されないが、チンパンジー、ボノボ、ゴリラ、ボルネオオランウータン、スマトラオランウータン、アカゲザル、及びマーモセットを含む、複数の非ヒト霊長類種からのミスセンス変異体を含む。

30

40

【0009】

本明細書で論じられるように、多数のプロセッサ上で実行される深層畳み込みニューラルネットワークは、変異体アミノ酸配列を良性又は病原性として分類するように訓練され得る。したがって、そのような深層畳み込みニューラルネットワークの出力は、変異体アミノ酸配列の病原性スコア又は分類を含み得るが、これらに限定されない。理解され得るように、特定の実装形態では、ニューラルネットワークベースの手法に加えて、又はその代わりに、訓練又は好適にパラメータ化された統計モデル若しくは技術及び／又は他の機械学習手法を用いることができる。

【0010】

50

本明細書で論じられる特定の実施形態では、病原性処理及び／又はスコアリング動作は更なる特徴又は態様を含み得る。例として、様々な病原性スコアリング閾値が、良性又は病原性として変異体を評価又は採点するためなど、評価又は査定プロセスの一部として用いられ得る。例として、特定の実装形態では、病原性である可能性の高い変異体のための閾値として使用するための遺伝子ごとの病原性スコアの好適なパーセンタイルは、51、55、65、70、75、80、85、90、95、又は99パーセンタイルなど、51%～99%の範囲内であり得るが、これらに限定されない。逆に、良性である可能性の高い変異体のための閾値として使用される遺伝子ごとの病原性スコアの好適なパーセンタイルは、限定されないが、1、5、10、15、20、25、30、35、40、又は45パーセンタイルなど、1%～49%の範囲内であり得る。

10

【0011】

更なる実施形態では、病原性処理及び／又はスコアリング動作は、選択効果を推定することを可能にする更なる特徴又は態様を含み得る。そのような実施形態では、変異率及び／又は選択を特徴付ける好適な入力を用いて、所与の集団内の対立遺伝子頻度の順方向時間シミュレーションが、関心のある遺伝子における対立遺伝子頻度スペクトルを生成するために使用され得る。その後、例えば、選択有り及び選択無しの対立遺伝子頻度スペクトルを比較することによって、関心のある変異体の枯渇メトリックが計算され得、対応する選択－枯渇関数が適合され又は特徴付けられ得る。所与の病原性スコア及びこの選択枯渇関数に基づいて、変異体について生成された病原性スコアに基づいて、所与の変異体について選択係数を求めることができる。

20

【0012】

追加の態様では、病原性処理及び／又はスコアリング動作は、病原性スコアを使用して遺伝的疾患の有病率を推定することを可能にする更なる特徴又は態様を含み得る。各遺伝子についての遺伝的疾患有病率メトリックを計算することに関して、第1の方法論では、病的変異体のセットのトリヌクレオチドコンテキスト構成が最初に取得される。このセット内の各トリヌクレオチドコンテキストについて、特定の選択係数（例えば、0.01）を仮定した順方向時間シミュレーションが実行され、そのトリヌクレオチドコンテキストの予想対立遺伝子頻度スペクトル（allele frequency spectrum、AFS）が生成される。遺伝子中のトリヌクレオチドの頻度によって重み付けされたトリヌクレオチドにわたるAFSを足し合わせることで、その遺伝子の予想AFSが生成される。この手法に係る遺伝的疾患有病率メトリックは、その遺伝子のための閾値を超える病原性スコアを有する変異体の予想累積対立遺伝子頻度として定義され得る。

30

【0013】

更なる態様では、病原性処理及び／又はスコアリング動作は、病原性スコアリングを再較正するための特徴又は方法論を含み得る。再較正などに関して、例示的な一実施形態では、再較正手法は、変異体の病原性スコアのパーセンタイルに焦点を当ててもよい。なぜなら、これらはより堅牢で、遺伝子全体に加えられる選択圧力による影響をあまり受けないことがあるためである。一実装形態によれば、病原性スコアの各パーセンタイルの生存確率が計算され、これは、病原性スコアのパーセンタイルが高いほど、変異体が浄化選択を生き残る可能性が低いことを示唆する生存確率補正係数を構成する。生存確率補正係数は、ミスセンス変異体における選択係数の推定に対するノイズの影響を軽減するのに役立つように再較正を実行するために用いられ得る。

40

【0014】

前述の説明は、開示される技術の作製及び使用を可能にするために提示されている。開示される実施態様に対する様々な修正は、明らかであり、本明細書で定義される一般原理は、開示される技術の趣旨及び範囲から逸脱することなく、他の実施態様及び用途に適用され得る。したがって、開示される技術は、示される実施態様に限定されることを意図するものではなく、本明細書に開示される原理及び特徴と一致する最も広い範囲を与えられるものである。開示される技術の範囲は、添付の特許請求の範囲によって規定される。

【図面の簡単な説明】

50

【0015】

本発明のこれらの特徴、態様、及び利点、並びに他の特徴、態様、及び利点は、添付図面を参照して以下の詳細な説明を読むと、より深く理解されることが考えられ、同様の特徴は、図面にわたって同様の部分を表している。

【図1】開示された技術の一実装形態に係る畳み込みニューラルネットワークを訓練する態様のブロック図を示す。

【図2】開示された技術の一実装形態に係る、タンパク質の二次構造及び溶媒露出度を予測するために使用される深層学習ネットワークアーキテクチャを示す。

【図3】開示された技術の一実装形態に係る病原性予測のための深層残差ネットワークの例示的なアーキテクチャを示す。

【図4】開示された技術の一実装形態に係る病原性スコア分布を示す。

【図5】開示された技術の一実装形態に係る、その遺伝子の全てのミスセンス変異体の75パーセンタイルにおける病原性スコアに対するClinVar病原性変異体の平均病原性スコアの相関関係のプロットを示す。

【図6】開示された技術の一実装形態に係る、その遺伝子の全てのミスセンス変異体の25パーセンタイルにおける病原性スコアに対するClinVar良性変異体の平均病原性スコアの相関関係のプロットを示す。

【図7】開示された技術の一実装形態に係る、それらの病原性スコアに基づいて、変異体を良性又は病原性のカテゴリに特徴付けるために閾値を使用し得るサンプルプロセスフローを示す。

【図8】開示された技術の一実装形態に係る、最適な順方向時間モデルパラメータが導出され得るサンプルプロセスフローを示す。

【図9】開示された技術の一実装形態に係る、異なる成長速度での指数関数的拡張の4つの段階に簡略化されたヒト集団の進化履歴を示す。

【図10】本手法及び他の文献から導出された変異率に従って導出された変異率の推定値間の相関関係を示す。

【図11】本開示の態様に係る、メチル化レベルに対するCpG変異の予想数の確認された比率を示す。

【図12A】本開示の態様に係る、順方向時間シミュレーションモデルの実装形態のための最適なパラメータ組み合わせを示すピアソンのカイ二乗統計量のヒートマップを示す。

【図12B】本開示の態様に係る、順方向時間シミュレーションモデルの実装形態のための最適なパラメータ組み合わせを示すピアソンのカイ二乗統計量のヒートマップを示す。

【図12C】本開示の態様に係る、順方向時間シミュレーションモデルの実装形態のための最適なパラメータ組み合わせを示すピアソンのカイ二乗統計量のヒートマップを示す。

【図12D】本開示の態様に係る、順方向時間シミュレーションモデルの実装形態のための最適なパラメータ組み合わせを示すピアソンのカイ二乗統計量のヒートマップを示す。

【図12E】本開示の態様に係る、順方向時間シミュレーションモデルの実装形態のための最適なパラメータ組み合わせを示すピアソンのカイ二乗統計量のヒートマップを示す。

【図13】一例では、本手法に従って決定された最適なモデルパラメータを使用して導出されたシミュレーション対立遺伝子頻度スペクトルは、確認された対立遺伝子頻度スペクトルに対応することを示す。

【図14】開示された技術の一実装形態に係る、順方向時間シミュレーションのコンテキストにおける選択効果が組み込まれるサンプルプロセスフローを示す。

【図15】本開示の態様に係る、選択-枯渇曲線の例を示す。

【図16】開示される技術の一実装形態に係る、関心のある変異体の選択係数が導出され得るサンプルプロセスフローを示す。

【図17】開示された技術の一実装形態に係る、病原性-枯渇関係が導出され得るサンプルプロセスフローを示す。

【図18】本開示の態様に係る、BRCA1遺伝子についての病原性スコア対枯渇のプロットを示す。

10

20

30

40

50

【図 19】本開示の態様に係る、LDLR 遺伝子についての病原性スコア対枯渇のプロットを示す。

【図 20】開示された技術の一実装形態に係る、累積対立遺伝子頻度が導出され得るサンプルプロセスフローを示す。

【図 21】開示された技術の一実装形態に係る、予想累積対立遺伝子頻度が導出され得る一般化されたサンプルプロセスフローを示す。

【図 22】本開示の態様に係る、予想累積対立遺伝子頻度と、確認された累積対立遺伝子頻度とのプロットを示す。

【図 23】本開示の態様に係る、予想累積対立遺伝子頻度と、疾患有病率とのプロットを示す。

【図 24】開示された技術の一実装形態に係る、予想累積対立遺伝子頻度が導出され得る第 1 のサンプルプロセスフローを示す。

【図 25】開示された技術の一実装形態に係る、予想累積対立遺伝子頻度が導出され得る第 2 のサンプルプロセスフローを示す。

【図 26】本開示の態様に係る、予想累積対立遺伝子頻度と、確認された累積対立遺伝子頻度とのプロットを示す。

【図 27】本開示の態様に係る、予想累積対立遺伝子頻度と、疾患有病率とのプロットを示す。

【図 28】開示された技術の一実装形態に係る、再校正手法の態様を病原性スコアリングプロセスの態様に関連付けるサンプルプロセスフローを示す。

【図 29】本開示の態様に係る、病原性スコアパーセンタイル対確率の分布を示す。

【図 30】本開示の態様に係る、ガウスノイズと重ね合わされた、確認された病原性スコアパーセンタイルの離散一様分布の密度プロットを示す。

【図 31】本開示の態様に係る、ガウスノイズと重ね合わされた、確認された病原性スコアパーセンタイルの離散一様分布の累積分布関数を示す。

【図 32】本開示の態様に係る、真の病原性スコアパーセンタイル (x 軸) を有する変異体が、確認された病原性スコアパーセンタイル間隔 (y 軸) に含まれる確率をヒートマップを介して示す。

【図 33】開示された技術の一実装形態に係る、補正係数の決定ステップのサンプルプロセスフローを示す。

【図 34】本開示の態様に係る、SCN2A 遺伝子のミスセンス変異体についての 10 本のピンのパーセンタイルにわたる枯渇確率を示す。

【図 35】本開示の態様に係る、SCN2A 遺伝子のミスセンス変異体についての 10 本のピンのパーセンタイルにわたる生存確率を示す。

【図 36】開示された技術の一実装形態に係る、補正された枯渇メトリックの決定ステップのサンプルプロセスフローを示す。

【図 37】本開示の態様に係る、真の病原性スコアパーセンタイル (x 軸) を有する変異体が、確認された病原性スコアパーセンタイル間隔 (y 軸) に含まれる確率を示す、補正又は再校正されたヒートマップを示す。

【図 38】本開示の態様に係る、各病原性スコアパーセンタイルピンについての補正された枯渇メトリックのプロットを示す。

【図 39】本開示の態様に係る、複数の層を有するフィードフォワードニューラルネットワークの一実装形態を示す。

【図 40】本開示の態様に係る、畳み込みニューラルネットワークの一実装形態の例を示す。

【図 41】本開示の態様に係る、特徴マップを追加することで事前情報を下流に再注入する残差接続を示す。

【図 42】開示された技術が動作可能なコンピューティング環境の例を示す。

【図 43】開示される技術を実装するために使用され得るコンピュータシステムの簡略ブロック図である。

10

20

30

40

50

【発明を実施するための形態】

【0016】

以下の考察は、開示される技術を当業者が作製及び使用することを可能にするために提示され、特定の用途及びその要件に関連して提供される。開示される実施態様に対する様々な修正は、当業者には容易に明らかとなり、本明細書で定義される一般原理は、開示される技術の趣旨及び範囲から逸脱することなく、他の実施態様及び用途に適用され得る。したがって、開示される技術は、示される実施態様に限定されることを意図するものではなく、本明細書に開示される原理及び特徴と一致する最も広い範囲を与えられるものである。

【0017】

I. 序論

以下の議論は、変異体病原性スコア又は分類の生成、及びそのような病原性スコア又は分類に基づく有用な臨床分析又はメトリックの導出など、以下で議論される特定の分析を実装するために使用され得る、畳み込みニューラルネットワークを含むニューラルネットワークの訓練及び使用に関連する態様を含む。これを念頭に置いて、本技術の説明においてそのようなニューラルネットワークの特定の態様及び特徴が言及又は参照され得る。議論を効率化するために、本技術の説明においてそのようなニューラルネットワークのベースライン知識が仮定される。しかし、関連するニューラルネットワーク概念の追加の説明を求める人達のために、記載の終盤において、関連するニューラルネットワーク概念の追加の情報及び説明が提供される。更に、有用な例を提供するために、及び説明を容易にするために本明細書ではニューラルネットワークが主に論じられているが、ニューラルネットワーク手法の代わりに又はニューラルネットワーク手法に加えて、訓練された若しくは好適にパラメータ化された統計モデル若しくは技術及び／又は他の機械学習手法を含むがこれらに限定されない他の実装形態が用いられ得ることも理解されたい。

【0018】

特に、以下の議論は、特定の関心のあるゲノムデータを分析するために使用される実装形態におけるニューラルネットワーク（例えば、畳み込みニューラルネットワーク）に関連する特定の概念を利用し得る。これを念頭に置いて、本明細書で説明されるニューラルネットワーク技術が利用され得る、本明細書で対処される問題に有用なコンテキストを提供するために、基礎となる関心のある生物学的問題及び遺伝的問題の特定の態様を概説する。

【0019】

遺伝的変異は多くの疾患を説明するのに役立ち得る。全てのヒトは固有の遺伝暗号を有し、個体の群内には多くの遺伝的変異体が存在する。病的な遺伝的変異体の多く又はほとんどは、自然選択によってゲノムから消失してきた。しかし、どの遺伝学的変異が病原性又は病的である可能性が高いかを特定することが依然として望ましい。特に、そのような知識は、研究者が病原性である可能性が高い遺伝的変異体に集中し、多くの疾患の診断及び治療のペースを加速するのに役立ち得る。

【0020】

変異体の特性及び機能的作用（例えば、病原性）をモデル化することは、ゲノムの分野における重要であるが困難なタスクである。機能的ゲノム配列決定技術の急速な進歩にもかかわらず、細胞型特異的転写調節システムの複雑さのために、変異体の機能的帰結の解釈は大きな課題である。したがって、変異体の病原性を予測するための強力な計算モデルは、基本科学及び翻訳研究の両方にとって大きな利益を有し得る。

【0021】

更に、過去数十年にわたる生化学的技術の進歩により、以前よりもはるかに低いコストでゲノムデータを迅速に生成する次世代配列決定（next generation sequencing、NGS）プラットフォームが誕生し、これは益々大量のゲノムデータを生成する。そのような非常に大量の配列決定されたDNAのアノテーションは依然として困難である。教師あり機械学習アルゴリズムは、通常、大量のラベル付きデータが利用可能である場合に十

10

20

30

40

50

分に機能する。しかしながら、バイオインフォマティクス及び多くの他のデータリッチ分野では、インスタンスのラベル付けプロセスは高コストである。逆に、ラベル無しインスタンスは安価で容易に入手可能である。ラベル付きデータの量が比較的少なく、ラベル無しデータの量が非常に多いシナリオの場合、半教師あり学習は、手動ラベル付けの費用効果の高い代替手段となる。したがって、これは、半教師ありアルゴリズムを使用して、変異体の病原性を正確に予測する深層学習ベースの病原性分類器を構築する機会を提示する。結果として、人間の確認バイアスを含まない病原性変異体のデータベースが得られ得る。

【0022】

機械学習ベースの病原性分類器に関して、深層ニューラルネットワークは、高レベル機能を連続的にモデル化するために、複数の非線形及び複雑な変換層を使用する、人工ニューラルネットワークの類である。深層ニューラルネットワークは、観測された出力と予測出力との間の差を伝達してパラメータを調整する逆伝搬を介してフィードバックを提供する。深層ニューラルネットワークは、大きな訓練データセットの利用可能性、並列分散コンピューティングの能力、及び高度な訓練アルゴリズムとともに進化している。

【0023】

畳み込みニューラルネットワーク (Convolutional neural network、CNN) 及び反復ニューラルネットワーク (Recurrent Neural Network、RNN) は、深層ニューラルネットワークの構成要素である。畳み込みニューラルネットワークは、畳み込み層、非線形層、及びプーリング層を含むアーキテクチャを有し得る。反復ニューラルネットワークは、パーセプトロン、長い短期メモリユニット、及びゲートされた反復単位のような構成単位間の周期的接続を有する入力データの連続的な情報を利用するように設計される。加えて、多くの他の新興の深層ニューラルネットワークが、深層時空間ニューラルネットワーク、多次元反復ニューラルネットワーク、及び畳み込み自動エンコーダなどの限定された状況に関して提案されてきた。

【0024】

配列データが多次元及び高次元であると仮定すると、深層ニューラルネットワークは、それらの広範な適用性及び強化された予測能力により、生物情報科学研究のために有望である。畳み込みニューラルネットワークは、モチーフ発見、病原性変異体識別、及び遺伝子発現推測などのゲノミクスにおける配列に基づく問題を解決するために採用されている。畳み込みニューラルネットワークは、DNAを研究するのに有用な重み共有戦略を使用するが、これは、短い配列モチーフを捕捉することができ、この配列モチーフは、有意な生物学的機能を有すると推定されるDNA中の局所的パターンを再現する。畳み込みニューラルネットワークの顕著な特徴は、畳み込みフィルタの使用である。精巧に設計され、手動で巧妙に作り上げられた特徴に基づく従来の分類アプローチとは異なり、畳み込みフィルタは、知識の情報表現に生入力データをマッピングするプロセスに類似した特徴の適応学習を実行する。この意味では、畳み込みフィルタは、そのようなフィルタのセットが入力内の関連するパターンを認識し、訓練手順中にそれ自体を更新することができるため、一連のモチーフスキャナーとして機能する。反復ニューラルネットワークは、タンパク質又はDNA配列などの様々な長さの連続的データにおける長距離依存性をとらえることができる。

【0025】

図1に概略的に示されるように、深層ニューラルネットワークの訓練は、各層における重みパラメータの最適化を伴い、このことは、より好適な階層表現がデータから学習されるように、より単純な特徴を複雑な特徴に徐々に組み合わせる。最適化プロセスの単一サイクルは、以下のように構成される。まず、訓練データセット (例えば、この例では入力データ100) と仮定すると、フォワードパスは、各層内の出力を順次計算し、ニューラルネットワーク102を通して機能信号を順方向に伝搬する。最終出力層において、客観的な損失関数 (比較ステップ106) は、推測された出力110と所与のラベル112との間の誤差104を測定する。訓練誤差を最小化するために、バックワードパスは、連

10

20

30

40

50

鎖ルールを使用して、誤差信号を逆伝搬し（ステップ114）、ニューラルネットワーク102全体の全ての重みに対する勾配を計算する。最後に、確率勾配降下に基づく最適化アルゴリズム、又は他の適切な手法を使用して重みパラメータが更新される（ステップ120）。バッチ勾配降下が完全データセットごとにパラメータ更新するのに対し、確率的勾配降下は、データ例の各々の小さいセットについて更新を実行することによって確率的近似値を提供する。いくつかの最適化アルゴリズムは確率的勾配降下に由来する。例えば、Adagrad and Adam訓練アルゴリズムは、それぞれ、各パラメータの更新頻度及び勾配のモーメントに基づいて学習率を適応的に修正しながら、確率的勾配降下を実行する。

【0026】

深層ニューラルネットワークの訓練における別の要素は規則化であり、規則化は、過剰適合を回避し、したがって良好な一般化性能を達成することを意図する戦略を指す。例えば、重み減衰は、重みパラメータがより小さい絶対値に収束するように、客観的損失関数にペナルティ項を追加する。ドロップアウトは、訓練中にニューラルネットワークから隠れたユニットをランダムに除去し、可能なサブネットワークの集合体と見なすことができる。更に、バッチ正規化は、ミニバッチ内の各アクティブ化に関するスカラー特徴の正規化を介した新たな規則化方法を提供し、各々の平均及び分散をパラメータとして学習する。

【0027】

ここで記載されている技術に関して、先行する高レベルの概要を考慮すると、ここで記載されている技術は、多数の人間によって操作された特徴及びメタ分類器を使用する従来の病原性分類モデルとは異なる。対照的に、本明細書に記載の技術の特定の実施形態では、関心のある変異体がフランキングしているアミノ酸配列と、他の種のオルソログ配列アラインメントとだけを入力として取得する単純な深層学習残差ネットワークを用いることができる。特定の実装形態では、タンパク質構造に関する情報をネットワークに提供するために、2つの別個のネットワークが、配列のみから二次構造及び溶媒露出度をそれぞれ学習するように訓練され得る。これらは、タンパク質構造に対する作用を予測するために、より大きな深層学習ネットワーク内のサブネットワークとして組み込まれ得る。出発点として配列を使用することにより、不完全に確認又は一貫して適用されないことがある、タンパク質構造及び機能ドメインアノテーションにおける潜在的なバイアスが回避される。

【0028】

深層学習分類器の精度は訓練データセットのサイズに対応し、6つの霊長類種の各々からの変異データは、分類器の精度の向上に独立して寄与する。現存するヒト以外の霊長類種の数の多さ及び多様性に加えて、タンパク質変異体に対する選択的圧力が霊長類系統内でほとんど一致していることを示す証拠は、臨床的ゲノム解釈を現在制限する未知の有意な数百万のヒト変異体を分類するための有効な戦略として、体系的な霊長類集団配列決定を示唆する。

【0029】

更に、コモン霊長類変異は、メタ分類器の増加のために客観的に評価することが困難であった、過去に使用された訓練データから完全に独立した既存の方法を評価するためのクリーンな検証データセットを提供する。10,000個の提出された霊長類コモン変異体を使用して、本明細書に記載の本モデルの性能を、4つの他のよく知られた分類アルゴリズム（Sift、Polyphen2、CADD、M-CAP）とともに評価した。全てのヒトミスセンス変異体のうちほぼ50%は、コモン対立遺伝子頻度で自然選択によって除去されるため、変異率によって10,000個の提出された霊長類コモン変異体にマッチされたランダム選択ミスセンス変異体のセットについて、分類器ごとに50パーセントイルスコアが計算され、その閾値を使用して、提出された霊長類コモン変異体が評価された。現在開示される深層学習モデルの精度は、ヒトのコモン変異体でのみ訓練された、又はヒトのコモン変異体及び霊長類変異体の両方を使用して訓練された深層学習ネットワークを使用するこの独立検証データセットについて、他の分類器よりも著しく良好であった

10

20

30

40

50

。

【0030】

上記に鑑み、要約すると、本明細書に記載の方法論は、変異体の病原性を予測するための既存の方法とは様々な形で異なる。最初に、現在記載されている手法は、半教師あり深層畳み込みニューラルネットワークの新規アーキテクチャを採用する。第二に、ヒトコモン変異体（例えば、gnomAD）及び霊長類変異体から信頼性の高い良性変異体が得られるとともに、バランスの取れたサンプリング及び訓練の反復を通じて信頼度が非常に高い病原性訓練セットが生成されることで、同一のキュレーションされたヒト変異体データベースを使用したモデルのサークル訓練及び試験が回避される。第三に、二次構造及び溶媒露出度のための深層学習モデルは、病原性モデルのアーキテクチャに統合される。構造モデル及び溶媒モデルから得られた情報は、特定のアミノ酸残基のためのラベル予測に限定されない。むしろ、構造モデル及び溶媒モデルから読み出し層が除去され、事前に訓練されたモデルが病原性モデルと融合される。病原性モデルを訓練する間、誤差を最小限に抑えるために構造及び溶媒の事前に訓練された層も逆伝搬する。これは、事前に訓練された構造及び溶媒モデルが病原性予測問題に集中するのを助ける。

10

【0031】

本明細書でも議論されているように、本明細書に記載されるように訓練及び使用されるモデルの出力（例えば、病原性スコア及び／又は分類）は、有用な追加データ又は診断、例えば、一連の臨床的に有意な変異体に対する選択効果の推定や遺伝的疾患有病率の推定などを生成するために使用され得る。モデル出力の再較正や、病原性変異体及び良性変異体の特徴付けるための閾値の生成及び使用など、他の関連する概念も説明される。

20

【0032】

II. 用語／定義

本明細書において、

塩基は、ヌクレオチド塩基又はヌクレオチド、（アデニン）、C（シトシン）、T（チミン）、又はG（グアニン）を指す。

【0033】

用語「タンパク質」及び「翻訳された配列」は同義で使用され得る。

【0034】

用語「コドン」及び「塩基トリプレット」は同義で使用され得る。

30

【0035】

用語「アミノ酸」及び「翻訳されたユニット」は同義で使用され得る。

【0036】

「変異体病原性分類器」、「変異体分類のための畳み込みニューラルネットワークベースの分類器」、及び「変異体分類のための深層畳み込みニューラルネットワークベースの分類器」という語句は同義で使用され得る。

【0037】

III. 病原性分類ニューラルネットワーク

A. 訓練及び入力

例示的な実装形態を参照すると、変異病原性分類（例えば、病原性又は良性）、及び／又は病原性若しくは病原性の欠如を数値的に特徴付ける定量的メトリック（例えば、病原性スコア）の生成に使用され得る深層学習ネットワークが本明細書において記載されている。良性ラベルを有する変異体のみを使用して分類器を訓練する1つのコンテキストでは、所与の変異が集団においてコモン変異体として確認される可能性が高いか否かとして予測問題が設定された。高い対立遺伝子頻度で変異体を確認される確率にはいくつかのファクタが影響を与えるが、本議論及び説明の主な焦点は有害性である。他のファクタとしては、変異率、配列決定カバレッジなどの技術的アーチファクト、及び中立遺伝子浮動（遺伝子変換など）に影響を与えるファクタが挙げられるが、これらに限定されない。

40

【0038】

深層学習ネットワークの訓練に関して、臨床用途での変異体分類の重要性は、教師あり

50

機械学習を使用して上記問題に対処する多数の試みを生じさせてきたが、これらの努力は、訓練のための高い信頼度でラベル付けされた良性及び病原性変異体を含む適切なサイズの真のデータセットの欠如によって妨げられてきた。

【0039】

人間の専門家によってキュレーションされた変異体の既存のデータベースは、ゲノム全体を表すものではなく、ClinVarデータベース中の変異体の約50%は、200個の遺伝子（ヒトタンパク質コード遺伝子の約1%）のみに由来する。更に、体系的研究により、多くの人間の専門家によるアノテーションが、疑わしい証拠によってサポートされていることが特定され、単一の患者のみで確認され得るレア変異体を解釈する困難性が強調される。人間の専門家による解釈は益々正確になっているが、分類ガイドラインは主にコンセンサス実践を中心に策定され、既存の傾向を強化するリスクがある。人間の解釈バイアスを低減するために、最近の分類器は、コモンヒト遺伝子多型又は固定ヒトチンパンジー置換で訓練されているが、これらの分類器も、人間によってキュレーションされたデータベースで訓練された以前の分類器の予測スコアを入力として使用する。これらの様々な方法の性能の客観的ベンチマーキングは、独立したバイアスのない真のデータセットが存在しない場合にはとらえどころがなかった。

【0040】

この問題に対処するために、現在説明される技術は、コモンヒト変異と重複せず、大半が浄化選択のふるいを通じた良性結果のコモン変異体を表す300,000個超のユニークなミスセンス変異体を与える、ヒトではない霊長類（例えば、チンパンジー、ボノボ、ゴリラ、オラウータン、及びマーモセット）からの変異を活用する。これにより、機械学習手法に利用可能な訓練データセットが大幅に拡大する。平均して、各霊長類種は、ClinVarデータベース全体（有意性が不確実な変異体及び矛盾するアノテーションを有するものを除くと、2017年11月時点で約42,000個のミスセンス変異体）よりも多くの変異体を与える。更に、このコンテンツはヒトの解釈のバイアスを含まない。

【0041】

本技術に従って使用される良性訓練データセットの生成に関して、あるそのようなデータセットは、大半が、機械学習のためのヒト及び非ヒト霊長類からのコモン良性ミスセンス変異体から構築された。データセットは、コモンヒト変異体（ $>0.1\%$ 対立遺伝子頻度、83,546個の変異体）と、チンパンジー、ボノボ、ゴリラ、及びオラウータン、アカゲザル、及びマーモセットからの変異体（301,690個のユニークな霊長類変異体）とを含んでいた。

【0042】

コモンヒト変異体（対立遺伝子頻度（AF） $>0.1\%$ ）及び霊長類変異を含むそのようなデータセットを使用して、本明細書ではPrimate AI又はpAIと称される深層残差ネットワークが訓練された。ネットワークは、関心のある変異体がランキングしているアミノ酸配列及び他の種のオルソログ配列アラインメントを入力として受け取るように訓練された。人間によって操作された特徴を使用する既存の分類器とは異なり、現在記載されている深層学習ネットワークは、一次配列から直接特徴を抽出するように訓練された。特定の実装形態では、タンパク質構造に関する情報を組み込むために、後に詳述されるように、配列のみからの二次構造及び溶媒露出度を予測するように別々のネットワークが訓練され、サブネットワークとしてフルモデル内に含まれた。十分に結晶化したヒトタンパク質の数は限られることから、一次配列から構造を推測することは、不完全なタンパク質構造及び機能ドメインアノテーションに起因するバイアスを回避するという利点を有する。タンパク質構造を含むネットワークの一実装形態の総深度は、約400,000個の訓練可能なパラメータを含む36層の畳み込みであった。

【0043】

B. タンパク質構造二次構造サブネットワーク及び溶媒露出度サブネットワーク

実装形態の一例では、病原性予測のための深層学習ネットワークは、二次構造サブネットワーク及び溶媒露出度予測サブネットワークのための19層の畳み込み層と、二次構造

サブネットワーク及び溶媒露出度サブネットワークの結果を入力として取るメイン病原性予測ネットワークのための17層とを含む、36の総畳み込み層を含む。特に、ほとんどのヒトタンパク質の結晶構造が不明であるため、二次構造ネットワーク及び溶媒露出度予測ネットワークは、ネットワークが一次配列からタンパク質構造を学習することを可能にするように訓練された。

【0044】

あるそのような実装形態における二次構造ネットワーク及び溶媒露出度予測ネットワークは同一のアーキテクチャ及び入力データを有するが、予測状態が異なる。例えば、あるそのような実装形態では、二次構造ネットワーク及び溶媒露出度ネットワークへの入力、ヒトと、99の他の脊椎動物との複数の配列アラインメントからの保全情報を符号化する好適な寸法（例えば、長さ51×20アミノ酸位置頻度行列（position frequency matrix、PFM））のアミノ酸PFMである。

【0045】

一実施形態では、図2を参照すると、病原性予測のための深層学習ネットワークと、二次構造及び溶媒露出度を予測するための深層学習ネットワークとが、残差ブロック140のアーキテクチャを採用した。残差ブロック140は繰り返しの畳み込みユニットを含み、先行する層からの情報が残差ブロック140をスキップすることを可能にするスキップ接続142が散りばめられている。各残差ブロック140において、入力層が最初にバッチ正規化され、次に、rectified linear unit (ReLU) を使用する活性化層が続く。次いで、活性化が1D畳み込み層を通過する。1D畳み込み層からのこの中間出力は再びバッチ正規化され、ReLU活性化され、別の1D畳み込み層へと続く。第2の1D畳み込みの終了時に、出力が残差ブロックへの元の入力と足し合わされる（ステップ146）。これは、元の入力情報が残差ブロック140をバイパスすることによってスキップ接続142として機能する。深層残差学習ネットワークと称され得るそのようなアーキテクチャでは、入力はその元の状態で保存され、残差接続は、モデルからの非線形活性化を含まず、より深いネットワークの効果的な訓練が可能になる。二次構造ネットワーク130及び溶媒露出度ネットワーク132の両方のコンテキストにおける詳細なアーキテクチャが図2、（後述される）表1及び表2に提供されている。ここで、PWM保全データ150が初期入力として示されている。図示の例では、モデルへの入力150は、（タンパク質構造データバンク配列上での訓練のために）RapptorXソフトウェア、又は（ヒトタンパク質配列上での訓練及び推測のための）99個の脊椎動物アラインメントによって生成された保全を使用した位置重み行列（position-weighted matrix、PWM）であり得る。

【0046】

残差ブロック140に続いて、ソフトマックス層154が、各アミノ酸について3つの状態の確率を計算し、その中で最大のソフトマックス確率がアミノ酸の状態を決定する。あるそのような実装形態のモデルは、ADAMオプティマイザを使用して、タンパク質配列全体についての累積カテゴリ交差エントロピー損失関数を用いて訓練される。例示される一実施形態では、ネットワークが二次構造及び溶媒露出度に関して事前に訓練された後、ネットワークの出力を病原性予測ネットワーク160の入力として直接取得する代わりに、より多くの情報が病原性予測ネットワーク160を通過するよう、ソフトマックス層154の前の層が代わりに取得された。一例では、ソフトマックス層154の前の層の出力は、好適な長さのアミノ酸配列（例えば、長さは51アミノ酸）であり、病原性分類のための深層学習ネットワークの入力になる。

【0047】

これに鑑み、二次構造ネットワークは、3状態二次構造体、すなわち、（1）αヘリックス（H）、（2）βシート（B）、又は（3）コイル（C）を予測するように訓練される。溶媒露出度ネットワークは、3状態の溶媒露出度、すなわち、（1）埋没（B）、（2）中間（I）、又は（3）露出（E）を予測するように訓練される。上記のように、両ネットワークは入力150として一次配列のみを取得し、タンパク質構造データバンクの

既知の結晶構造からのラベルを使用して訓練された。各モデルモデルは、各アミノ酸残基についてそれぞれ1つの状態を予測する。

【0048】

上記に鑑み、例示的な実装形態の更なる例示として、入力データセット150内の各アミノ酸位置について、ランキングしているアミノ酸（例えば、ランキングしている51のアミノ酸）に対応して位置頻度行列から窓が取得され、これを使用して、長さアミノ酸配列の中心のアミノ酸についての二次構造又は溶媒露出度のためのラベルが予測された。二次構造及び相対溶媒露出度のラベルは、DSSPソフトウェアを使用して、タンパク質の既知の3D結晶構造から直接取得され、一次配列からの予測を必要としなかった。二次構造ネットワーク及び溶媒露出度ネットワークを病原性予測ネットワーク160の一部として組み込むために、ヒトベースの99個の脊椎動物マルチ配列アラインメントから位置頻度行列が計算された。これら2つの方法から生成された保全行列は一般的に同様であるが、パラメータ重みの微調整を可能にするために、病原性予測のための訓練中の二次構造モデル及び溶媒露出度モデルを介する逆伝搬が可能にされた。

10

【0049】

例として、表1は、3状態二次構造予測深層学習（deep learning、DL）モデルの例示的なモデルアーキテクチャの詳細を示す。形状は、モデルの各層における出力テンソルの形状を指定し、活性化は、層のニューロンに与えられる活性化である。モデルへの入力、変異体の周りのランキングしているアミノ酸配列のための好適な寸法（例えば、長さが51アミノ酸、深度が20）の位置特異的頻度行列であった。

20

【0050】

同様に、表2に示されるモデルアーキテクチャは、本明細書に記載されるように、二次構造予測DLモデルとのアーキテクチャが同一であり得る3状態溶媒露出度予測深層学習モデルの例示的なモデルアーキテクチャの詳細を示す。形状は、モデルの各層における出力テンソルの形状を指定し、活性化は、層のニューロンに与えられる活性化である。モデルへの入力、変異体の周りのランキングしているアミノ酸配列のための好適な寸法（例えば、長さが51アミノ酸、深度が20）の位置特異的周波数行列であった。

【0051】

3状態二次構造予測モデルの最高の試験精度は80.32%であった。これは、同様の訓練データセットを用いてDeepCNFモデルによって予測される最先端の精度と同様である。3状態溶媒露出度予測モデルの最高の試験精度は64.83%であった。これは、同様の訓練データセットを用いてRaptorXによって予測される現在の最高の精度と同様である。

30

【0052】

例示的な実装形態—モデルアーキテクチャ及び訓練

（以下に再現される）表1及び表2、並びに図2を参照して、かつ実装形態の例を提供することによって、それぞれタンパク質の3状態二次構造及び3状態溶媒露出度を予測するために、2つのエンドツーエンド深層畳み込みニューラルネットワークモデルが訓練された。2つのモデルは、2つの入力チャンネルを含む同様の構成を有し、一方はタンパク質配列のためのものであり、もう一方はタンパク質保全プロファイルのためのものである。各入力チャンネルは $L \times 20$ の寸法を有し、ここで、 L はタンパク質の長さを示す。

40

【0053】

各入力チャンネルが、40のカーネル及び線形活性化を有する1D畳み込み層（層1a及び1b）を通過させられた。この層を使用して、入力寸法が20から40にアップサンプリングされた。モデル全体を通して、全ての他の層が40カーネルを使用した。2つの層（1a及び1b）の活性化は、40の寸法の各々にわたる値を足し合わせるによって融合された（すなわち、マージモード＝「和」）。マージノードの出力は1D畳み込み（層2）の単一層を通過させられ、その後、線形活性化が行われた。

【0054】

層2からの活性化は、一連の9個の残差ブロック（層3～11）を通過させられた。層

50

3の活性化は層4に供給され、層4の活性化は層5に供給される、などであった。全ての3つおきの残差ブロック（層5、8、及び11）の出力を直接足し合わせるスキップ接続も存在した。次いで、融合された活性化は、ReLU活性化を有する2つの1D畳み込み（層12及び層13）に供給された。層13からの活性化は、ソフトマックス読み出し層に与えられた。ソフトマックスは、与えられた入力に対して3つのクラスの出力の確率を計算した。

【0055】

更に注目すべきことに、二次構造モデルの一実装形態では、1D畳み込みのアトラスレートは1であった。溶媒露出度モデルのある実装形態において、最後の3つの残差ブロック（層9、10、及び11）のアトラスレートは、カーネルのカバレッジを増加させるために2であった。これらの態様に関して、アトラス／膨張畳み込みは、カーネルが、アトラス畳み込みレート又は膨張係数とも呼ばれる特定のステップで入力値をスキップすることによって、長さよりも大きい領域にわたって適用される畳み込みである。アトラス／膨張畳み込みは、畳み込みフィルタ／カーネルの要素間に空間を追加し、結果として、畳み込み演算が実行されるとき、より大きな間隔の隣接する入力エントリ（例えば、ヌクレオチド、アミノ酸）が考慮される。これにより、入力における長距離コンテキスト依存性の組み込みが可能になる。アトラス畳み込みは、隣接するヌクレオチドが処理される際の再利用のために部分的な畳み込み計算を保全する。アトラス／膨張畳み込みは、少数の訓練可能なパラメータを有する大きな受容野を可能にする。タンパク質の二次構造は、近接するアミノ酸の相互作用に強く依存する。したがって、より高いカーネルカバレッジを有するモデルは、性能をわずかに改善した。逆に、溶媒露出度はアミノ酸間の長距離相互作用の影響を受ける。したがって、アトラス畳み込みを使用するカーネルカバレッジの高いモデルの場合、精度は短カバレッジモデルの精度よりも2%超高かった。

【0056】

10

20

30

40

50

【表 1】

表 1－3 状態二次構造予測モデルの例

層	タイプ	カーネルの数、 窓サイズ	形状	アトラス レート	活性化
配列を入力(層1a)	畳み込み1D	40, 1	(L, 40)	1	線形
PSSMを入力(層1b)	畳み込み1D	40, 1	(L, 40)	1	線形
配列+PSSMを融合	融合(モード=連結)	—	(L, 80)	—	—
層2	畳み込み1D	40, 5	(L, 40)	1	線形
層3	畳み込み1D	40, 5	(L, 40)	1	ReLU
層4	畳み込み1D	40, 5	(L, 40)	1	ReLU
層5	畳み込み1D	40, 5	(L, 40)	1	ReLU
層6	畳み込み1D	40, 5	(L, 40)	1	ReLU
層7	畳み込み1D	40, 5	(L, 40)	1	ReLU
層8	畳み込み1D	40, 5	(L, 40)	1	ReLU
層9	畳み込み1D	40, 5	(L, 40)	1	ReLU
層10	畳み込み1D	40, 5	(L, 40)	1	ReLU
層11	畳み込み1D	40, 5	(L, 40)	1	ReLU
活性化を融合	融合—層5、8、及び 11。モード=「和」	—	(L, 40)	—	—
層12	畳み込み1D	40, 5	(L, 40)	1	ReLU
層13	畳み込み1D	40, 5	(L, 40)	1	ReLU
出力層	畳み込み1D	1, 3	(L, 3)	—	ソフトマックス

【 0 0 5 7 】

10

20

30

40

50

【表 2】

表 2－3 状態溶媒露出度モデルの例

層	タイプ	カーネルの数、 窓サイズ	形状	アトラス レート	活性化
入力配列(層1a)	畳み込み1D	40, 1	(L, 40)	1	線形
入力PSSM(層1b)	畳み込み1D	40, 1	(L, 40)	1	線形
配列+PSSMを融合	融合(モード=連結)	—	(L, 80)	—	—
層2	畳み込み1D	40, 5	(L, 40)	1	線形
層3	畳み込み1D	40, 5	(L, 40)	1	ReLU
層4	畳み込み1D	40, 5	(L, 40)	1	ReLU
層5	畳み込み1D	40, 5	(L, 40)	1	ReLU
層6	畳み込み1D	40, 5	(L, 40)	1	ReLU
層7	畳み込み1D	40, 5	(L, 40)	1	ReLU
層8	畳み込み1D	40, 5	(L, 40)	1	ReLU
層9	畳み込み1D	40, 5	(L, 40)	2	ReLU
層10	畳み込み1D	40, 5	(L, 40)	2	ReLU
層11	畳み込み1D	40, 5	(L, 40)	2	ReLU
活性化を融合	融合—層5、8、及び 11。モード=「和」	—	(L, 40)	—	—
層12	畳み込み1D	40, 5	(L, 40)	1	ReLU
層13	畳み込み1D	40, 5	(L, 40)	1	ReLU
出力層	畳み込み1D	1, 3	(L, 3)	—	ソフトマックス

【0058】

C. 病原性予測ネットワークアーキテクチャ

病原性予測モデルに関して、変異体の病原性を予測するために半教師あり深層畳み込みニューラルネットワーク(CNN)モデルが開発された。モデルに入力される特徴には、変異体がフランキングしているタンパク質配列及び保全プロファイル、並びに特定の遺伝子領域におけるミスセンス変異体の枯渇が含まれる。変異体によって引き起こされる二次構造及び溶媒露出度への変化も深層学習モデルによって予測され、病原性予測モデルに統合された。あるそのような実装形態における予測病原性は、0(良性)～1(病原性)のスケールを有する。

【0059】

あるそのような病原性分類ニューラルネットワーク(例えば、PrimateAI)のアーキテクチャは図3に概略的に記載されており、(後述される)表3ではより詳細に記載されている。図3に示される例では、1Dは一次元畳み込み層を指す。他の実装形態では、モデルは、2D畳み込み、3D畳み込み、膨張若しくはアトラス畳み込み、転置畳み込み、分離可能畳み込み、及び深さ方向(depthwise)分離可能畳み込みなどの異なるタイプの畳み込みを使用することができる。更に、上記のように、病原性予測(例えば、PrimateAI又はpAI)のための深層学習ネットワークと、二次構造及び溶媒露出度を予測するための深層学習ネットワークとの両方の特定の实装形態は、残差ブロック

のアーキテクチャを採用した。

【0060】

特定の実施形態では、深層残差ネットワークのいくつか又は全ての層はReLU活性化関数を使用し、これは、シグモイド又は双曲線正接などの飽和非線形性と比較して、確率的勾配降下の収束を大幅に加速させる。開示される技術によって使用され得る活性化関数の他の例には、パラメトリックReLU、漏れReLU、及びELU (exponential linear unit) が含まれる。

【0061】

本明細書に記載されるように、いくつか又は全ての層はまた、バッチ正規化を使用して、訓練中に畳み込みニューラルネットワーク (CNN) 内の各層の分布を変化させてもよく、これは、層ごとに異なる。これは、最適化アルゴリズムの収束速度を下げる。

【0062】

10

20

30

40

50

【表 3】

表 3 ー一実施形態に係る病原性予測モデル

層	タイプ	カーネルの数、 窓サイズ	形状	アトラス レート	活性化
参照配列 (1a)	畳み込み 1D	40, 1	(51, 40)	1	線形
代替配列 (1b)	畳み込み 1D	40, 1	(51, 40)	1	線形
霊長類保全 (1c)	畳み込み 1D	40, 1	(51, 40)	1	線形
哺乳類保全 (1d)	畳み込み 1D	40, 1	(51, 40)	1	線形
脊椎動物保全 (1e)	畳み込み 1D	40, 1	(51, 40)	1	線形
参照配列ベースの二 次構造 (1f)	入力層	—	(51, 40)	—	—
代替配列ベースの二 次構造 (1g)	入力層	—	(51, 40)	—	—
参照配列ベースの溶 媒露出度 (1h)	入力層	—	(51, 40)	—	—
代替配列ベースの溶 媒露出度 (1i)	入力層	—	(51, 40)	—	—
参照配列融合 (2a)	融合 (モード=和) (1a, 1c, 1d, 1e)	—	(51, 40)	—	—
代替配列融合 (2b)	融合 (モード=和) (1b, 1c, 1d, 1e)	—	(51, 40)	—	—
3a	畳み込み 1D (2a)	40, 5	(51, 40)	1	ReLU
3b	畳み込み 1D (2b)	40, 5	(51, 40)	1	ReLU
4	融合 (モード=連結) (3a, 3b, 1f, 1g, 1h, 1i)	—	(51, 240)	—	—
5	畳み込み 1D	40, 5	(51, 40)	1	ReLU
6	畳み込み 1D	40, 5	(51, 40)	1	ReLU
7	畳み込み 1D	40, 5	(51, 40)	1	ReLU
8	畳み込み 1D	40, 5	(51, 40)	1	ReLU
9	畳み込み 1D	40, 5	(51, 40)	2	ReLU
10	畳み込み 1D	40, 5	(51, 40)	2	ReLU
11	畳み込み 1D	40, 1	(51, 1)	2	シグモイド
12	グローバル最大プー リング	—	1	—	—
出力層	活性化層	—	1	—	シグモイド

【0063】

例示的な実装形態ーモデルアーキテクチャ

上記を念頭に、図 3 及び表 3 を参照して、一実装形態では、病原性予測ネットワークは 5 つの直接入力及び 4 つの間接入力を受信する。このような例における 5 つの直接入力は、好適な寸法のアミノ酸配列（例えば、長さが 51 であるアミノ酸配列×深さ 20）（20 個の異なるアミノ酸を符号化）を含み得、変異体を有さない参照ヒトアミノ酸配列（1a）、置換変異体を有する代替ヒトアミノ酸配列（1b）、霊長類種の複数の配列アラインメントからの位置特異的頻度行列（PFM）（1c）、哺乳動物種の複数の配列アラインメントからの PFM（1d）、及びより離れた脊椎動物種（1e）の複数の配列アライ

ンメントからの P F M を備え得る。間接入力は、参照配列ベースの二次構造（1 f）、代替配列ベースの二次構造（1 g）、参照配列ベースの溶媒露出度（1 h）、及び代替配列ベースの溶媒露出度（1 i）を含む。

【0064】

間接入力 1 f 及び 1 g に関して、ソフトマックス層を除き、二次構造予測モデルの事前に訓練された層がロードされる。入力 1 f に関して、事前に訓練された層は、変異体のためのヒト参照配列、及び変異体のための P S I - B L A S T によって生成された P S S M に基づいたものである。同様に、入力 1 g に関して、二次構造予測モデルの事前に訓練された層は、P S S M 行列とともに入力としてのヒト代替配列に基づいたものである。入力 1 h 及び 1 i は、それぞれ、変異体の参照配列及び変異体の代替配列のための溶媒露出度

10

【0065】

この例では、5つの直接入力チャンネルが、線形活性化を有する 40 カーネルのアップサンプリング畳み込み層を通過する。層 1 a、1 c、1 d、及び 1 e は、40 の特徴寸法にわたって足し合わせられた値と融合され、層 2 a を生成する。言い換えれば、参照配列の特徴マップは 3 つのタイプの保全特徴マップと融合される。同様に、1 b、1 c、1 d、及び 1 e は、40 の特徴寸法にわたって足し合わせられた値と融合されて層 2 b を生成し、すなわち、代替配列の特徴は 3 つのタイプの保存特徴と融合される。

【0066】

層 2 a 及び 2 b は R e L U の活性化を用いてバッチ正規化され、各々がフィルタサイズ 40 の 1 D 畳み込み層（3 a 及び 3 b）を通過する。層 3 a 及び 3 b の出力は 1 f、1 g、1 h、及び 1 i と融合され、特徴マップが互いに連結される。言い換えれば、保全プロファイルを有する参照配列の特徴マップ、及び保全プロファイルを有する代替配列の特徴マップが、参照及び代替配列の二次構造特徴マップ並びに参照及び代替配列の溶媒露出度特徴マップと融合される（層 4）。

20

【0067】

層 4 の出力は、6 つの残差ブロック（層 5、6、7、8、9、10）を通過する。最後の 3 つの残差ブロックは、より高いカーネルカバレッジを提供するために、1 D 畳み込みについて 2 のアトラスレートを有する。層 10 の出力は、フィルタサイズ 1 の 1 D 畳み込み及び活性化シグモイド（層 11）を通過する。層 11 の出力は、変異体の単一の値を得るグローバル最大プーリングを通過する。この値は変異体の病原性を表す。病原性予測モデルの一実施の詳細を表 3 に示す。

30

【0068】

D. 訓練（半教師あり）及びデータ分布

半教師あり学習手法に関して、そのような技術は、ネットワーク（複数可）を訓練するためにラベル付きデータ及びラベル無しデータの両方の利用を許容する。半教師あり学習を選択する動機は、人間によってキュレーションされた変異体データベースが信頼できず、ノイズが多いこと、特に、信頼できる病原性変異体を欠いていることである。半教師あり学習アルゴリズムは訓練プロセスにおいてラベル付きインスタンス及びラベル無しインスタンスの両方を使用するため、訓練に利用可能な少量のラベル付きデータしか有さない完全に教師ありの学習アルゴリズムよりも優れた性能を達成する分類器を生成することができる。半教師あり学習の背後にある原理は、ラベル付きインスタンスのみを使用する教師ありモデルの予測能力を強化するために、ラベル無しデータに内在する知識を活用することであり、それによって、半教師あり学習の潜在的な利点が提供される。少量のラベル付きデータから教師あり分類器によって学習されたモデルパラメータは、ラベル無しデータによって、より現実的な（試験データの分布により似ている）分布に向かい得る。

40

【0069】

バイオインフォマティクスで一般的な別の課題は、データ不均衡問題である。データ不均衡現象は、予測されるべきクラスのうちの 1 つのクラスに属するインスタンスがレアである（注目すべき事例）か又は取得が困難であるため、そのクラスがデータにおいて不十

50

分である場合に生じる。マイノリティクラスは、典型的には、特別な事例に関連付けられ得ることから学習する重要性が最も高い。

【0070】

不均衡なデータ分布を処理するためのあるアルゴリズム手法は、分類器のアンサンブルに基づく。ラベル付きデータの限られた量は自然とより弱い分類器につながるが、弱い分類器のアンサンブルは、任意の単一の構成要素分類器の性能を上回る傾向がある。更に、アンサンブルは、典型的には、複数のモデルを学習することに関連付けられた労力及びコストを検証するファクタによって、単一の分類器から得られる予測精度を改善する。直感的には、個々の分類器の大きな多様性を平均化すると、分類器の過剰適合も平均化することになるため、いくつかの分類器を集合化することは、より良好な過剰適合管理につながる。

10

【0071】

IV. 遺伝子特異的病原性スコア閾値

ニューラルネットワークとして実装される病原性分類器の訓練及び検証に関する記載をしてきたが、以下の項は、そのようなネットワークを使用して病原性分類を更に改良及び／又は利用するための様々な実装形態固有シナリオ及び使用事例シナリオに関する。第1の態様では、スコアリング閾値及びそのようなスコア閾値の使用の議論が記載される。

【0072】

本明細書に記載されるように、本明細書に記載のPrimate AI又はpAI分類器などの現在開示されている病原性分類ネットワークは、遺伝子内の良性変異体から病原性変異体を区別又はスクリーニングするのに有用な病原性スコアを生成するために使用され得る。本明細書に記載の病原性スコアリングは、ヒト及びヒト以外の霊長類における浄化選択の程度に基づいているため、強力な浄化選択下にある遺伝子では、病原性変異体及び良性変異体に関連付けられた病原性スコアはより高いと予想される。一方、中立進化又は弱い選択下の遺伝子の場合、病原性変異体の病原性スコアはより低くなる傾向がある。この概念は図4に視覚的に示されており、対応する遺伝子のスコアの分布において、ある変異体の病原性スコア206が示されている。図4を参照して理解され得るように、実際には、病原性又は良性である可能性が高い変異体を同定するための近似遺伝子固有閾値を有することが有用であり得る。

20

【0073】

病原性スコアを評価するのに有用であり得る閾値を評価するために、ClinVarにおいて少なくとも10の良性／良性である可能性の高い変異体と、少なくとも10の病原性及び病原性である可能性が高い変異体とを含む84個の遺伝子を使用して、潜在的なスコア閾値を研究した。これらの遺伝子は、各遺伝子のための好適な病原性スコア閾値を評価するのに補助するために使用された。これらの遺伝子の各々について、ClinVar内の良性変異体及び病原性変異体について平均病原性スコアを測定した。

30

【0074】

一実装形態では、病原性変異体の遺伝子特異的Primate AI閾値を示す図5、及び良性変異体の遺伝子特異的Primate AI閾値を示す図6にグラフで示されるように、各遺伝子における病原性ClinVar変異体及び良性ClinVar変異体の平均病原性スコアが、その遺伝子における75パーセンタイル及び25パーセンタイルの病原性スコア（ここではPrimate AI又はpAIスコア）と十分に相関していることが確認された。両方の図において、各遺伝子はドットによって表されており、上方には遺伝子記号ラベルがある。これらの例では、ClinVar病原性変異体の平均Primate AIスコアは、その遺伝子の全てのミスセンス変異体の75パーセンタイルのPrimate AIスコアと密接に相関している（スピアマン相関=0.8521、図5）。同様に、ClinVar良性変異体の平均Primate AIスコアは、その遺伝子の全てのミスセンス変異体の25パーセンタイルのPrimate AIスコアと密接に相関している（スピアマン相関=0.8703、図6）。

40

【0075】

50

本手法を考慮すると、病原性である可能性が高い変異体のカットオフとして使用される遺伝子ごとの病原性スコアの好適なパーセンタイルは、51パーセンタイル以上かつ99パーセンタイル以下によって定義される範囲内であり得る（例えば、65、70、75、80、又は85パーセンタイル）。逆に、良性である可能性が高い変異体のカットオフとして使用される遺伝子ごとの病原性スコアの好適なパーセンタイルは、1パーセンタイル以上かつ49パーセンタイル以下によって定義される範囲内であり得る（例えば、15、20、25、30、又は35パーセンタイル）。

【0076】

これらの閾値の使用に関して、図7は、そのような閾値を使用して、変異体の病原性スコア206に基づいて変異体を良性又は病原性のカテゴリに分類することができるサンプルプロセスフローを示す。この例では、関心のある変異体200は、本明細書に記載されるような病原性スコアリングニューラルネットワークを使用して処理され得（ステップ202）、関心のある変異体200の病原性スコア206が導出され得る。図示の例では、病原性スコアは、遺伝子固有の病原性閾値212（例えば、（75%）と比較され（判定ブロック210）、病原性であると判定されない場合、遺伝子固有の良性閾値218と比較される（判定ブロック216）。この例における比較プロセスは、単純にするために直列に発生するものとして示されているが、実際には、比較は、単一のステップで、並列して、又は1つの比較のみが択一的に実行されてもよい（例えば、変異体が病原性であるか否かを判定する）。病原性閾値212を超える場合、関心のある変異体200は病原性変異体220と見なされ得る。逆に、病原性スコア206が良性閾値212未満である場合、関心のある変異体200は良性変異体222と見なされ得る。どちらの閾値基準も満たさない場合、関心のある変異体は病原性でも良性でもないものとして扱われ得る。ある研究において、本明細書に記載の手法を使用して、ClinVarデータベース内の17,948のユニークな遺伝子について遺伝子固有閾値及びメトリックが導出及び評価された。

10

20

【0077】

V. 順方向時間シミュレーションを使用した病原性スコアに基づく全てのヒト変異体に対する選択効果の推定

臨床研究及び患者の治療は、PrimateAIなどの病原性分類ネットワークが、遺伝子内の良性変異体から病原性変異体を分類及び／又は分けるために用いられ得る例示的な使用事例シナリオである。特に、臨床ゲノム配列決定は、レアな遺伝的疾患を有する患者のための標準治療となってきた。レアな遺伝的疾患は、多くの場合、あるいはほとんどの場合、非常に病的なレア変異によって引き起こされ、これらは一般的に、これらの変異の深刻さのために検出がより容易である。しかし、コモンな遺伝的疾患の根底にあるレアな変異は、それらの変異の弱い影響及び大きな数のためにほとんど特徴付けられていないままである。

30

【0078】

これに鑑み、特に本明細書で論じられるような変異体の病原性スコアリングのコンテキストにおいて、レアな変異とコモンな疾患との間のメカニズムを理解し、ヒト変異の進化ダイナミクスを研究することが望ましい可能性がある。ヒト集団の進化の間、デノボ変異によって新しい変異体が常に生成されてきた。それらのうちの一部は自然選択のために除去されてきた。ヒト集団サイズが一定であった場合、2つの力による影響を受ける変異体の対立遺伝子頻度は、最終的に平衡に達するであろう。これに鑑み、確認された対立遺伝子頻度を使用して、任意の変異体に対する浄化選択の深刻さを決定することが望ましい可能性がある。

40

【0079】

しかし、ヒト集団はいかなる時点でも定常状態ではなく、農業の出現から指数関数的に拡大している。したがって、本明細書で論じられる特定の手法によれば、変異体の対立遺伝子頻度の分布に対する2つの力の影響を調査するために、順方向時間シミュレーションがツールとして使用され得る。この手法の態様は、図8に示されるステップに関連して説

50

明される。図 8 は、最適な順方向時間モデルパラメータ導出の説明に際し参照されたり、再参照され得る。

【0080】

これに鑑み、デノボ変異率 280 を使用した中立進化の順方向時間シミュレーションが、経時的な変異体の対立遺伝子頻度分布のモデリングの一部として用いられ得る（ステップ 282）。ベースラインとして、順方向時間集団モデルは中立進化を仮定してシミュレートされ得る。モデルパラメータ 300 は、シミュレートされた対立遺伝子頻度スペクトル（AFS）304 を、ヒトゲノムにおいて確認された同義変異（同義 AFS 308）のそれに適合させることによって導出された（ステップ 302）。最適なモデルパラメータ 300 のセット（すなわち、最良の適合に対応するパラメータ）を使用して生成されたシミュレートされた AFS 304 が、有用な臨床情報を導出するために、変異体病原性スコアリングなどの本明細書で論じられる他の概念とあわせて使用され得る。

【0081】

レア変異体の分布が主な関心対象であるため、単純化されたヒト集団拡張モデル（すなわち、単純化された進化の歴史 278）の概略図である図 9 に示すように、この例示的失踪携帯におけるヒト集団の進化の歴史が、このシミュレーションでは異なる成長速度を有する 4 つの指数関数的拡張ステージに単純化されている。この例では、センサス集団サイズと有効集団サイズとの間の比率は、 r として示され得、初期有効集団サイズは $N_{e0} = 10,000$ として示され得る。各世代が約 30 年かかると仮定され得る。

【0082】

この例では、有効集団サイズの変化が小さい、長い $burn-in$ 期間（約 3,500 世代）を第 1 のフェーズで用いた。集団サイズの変化は n と示され得る。 $burn-in$ 後の時間は不明であることから、この時間は $T1$ と示され得、 $T1$ における有効集団サイズは $10,000 * n$ と示され得る。 $burn-in$ 中の成長速度 284 は $g1 = n^{(1/3,500)}$ である。

【0083】

1400 年において、世界のセンサス集団サイズは約 3 億 6,000 万であったと推定されている。1700 年において、センサス集団サイズは約 6 億 2,000 万まで増加し、2000 年では 62 億である。これらの推定値に基づいて、各ステージでの成長速度 284 は表 4 に示すように導出することができる。

【0084】

【表 4】

表 4：シミュレーションスキーム

時点	経過世代数	有効集団サイズ	センサス集団サイズ	成長速度
T1	大きな $burn-in$ (例えば、3500)	$10,000 * n$	$10,000 * r * n$	$n^{(1/3500)}$
T2:1400AD	$(T2 - T1) / 30$	$36000 / r$	360m	$(3.6 / r / n)^{((T2 - T1) / 30)}$
T3:1700AD	10	$62000 / r$	620m	$(6.2 / 3.6)^{(1/10)}$
T4:2000AD	10	$620000 / r$	6.2b	$10^{(1/10)}$

【0085】

世代 j 286 について、新世代集団を形成するために前の世代から N_j の染色体がランダムにサンプリングされた。ここで、 $N_j = g_j * N_{j-1}$ であり、 g_j は世代 j における成長速度 284 である。変異のほとんどは、染色体サンプリング中に前の世代から引き継がれる。デノボ変異率 (μ) 280 に従い、これらの染色体にデノボ変異が適用される。

10

20

30

40

50

【0086】

デノボ変異率280は、特定の実装形態によれば、以下の手法又は同等の手法に従って導出され得る。特に、あるそのような実装形態では、文献ソース（Hall d o r s s o n セット（2976トリオ）、G o l d m a n n セット（1291トリオ）、及び S a n d e r s セット（3804トリオ）からの全ゲノム配列決定を用いて、3つの大きな親子トリオデータセット（総数8,071トリオ）が取得された。これら8,071トリオを融合して、遺伝子間領域にマッピングされたデノボ変異が取得され、192のトリヌクレオチドコンテキスト構成の各々についてデノボ変異率280が導出された。

【0087】

図10に示すように、これらの変異率推定を他の文献の変異率（1,000のゲノムプロジェクトの遺伝子間領域から導出された K a i t l i n の変異率）と比較した。相関は0.9991であり、表5に示されるように、推定値は概ね K a i t l i n の変異率よりも低かった（ここで、C p G T i = C p G サイトでのトランジション変異、非 C p G T i = 非 C p G サイトでのトランジション変異、T v = トランスバージョン変異）。

【0088】

【表5】

表5－変異率の比較

変異率	Kaitlinの変異率			デノボ変異		
	加重平均	平均	中央値	加重平均	平均	中央値
ゲノムワイド	1.2e-8			1.048189e-08		
非CpGTi (56trinucs)	6.634517e-09	6.700049e-09	6.449612e-09	5.493724e-09	5.552238e-09	5.339421e-09
Tv	2.101242e-09	2.345956e-09	1.818607e-09	1.922048e-09	2.034197e-09	1.663312e-09
CpGTi	9.33749e-08	9.486598e-08	8.917675e-08	9.040611e-08	9.174335e-08	8.593169e-08
CpGTi(高メチル化)					1.011751e-07	
CpGTi(低メチル化)					2.264356e-08	

【0089】

C p G アイランドにおける変異率に関して、C p G サイトでのメチル化レベルは、実質的に変異率に影響を及ぼす。C p G T i 変異率を正確に計算するには、それらのサイトでのメチル化レベルを考慮すべきである。これに鑑み、例示的な実現形態では、変異率及び C p G アイランドは以下の手法に従って計算され得る。

【0090】

最初に、C p G 変異率に対するメチル化レベルの影響を、（R o a d m a p E p i g e n o m i c s プロジェクトから取得された）全ゲノムバイサルファイト配列決定データを使用することによって評価した。各 C p G アイランドのメチル化データを抽出し、10個の胚性幹細胞（E S C）試料全体で平均化した。次いで、図11に示すように、10個の定義されたメチル化レベルに基づいて、それらの C p G アイランドを10本のビンに分けた。各メチル化ビンに含まれる C p G サイトの数、及び C p G トランジション変異体の確認された数を遺伝子間領域及びエクソン領域の両方においてそれぞれカウントした。C p G T i 変異体の総数に、当該メチル化ビン内の C p G サイトの割合を乗じたものを、各メチル化ビンにおける C p G サイトでのトランジション変異体の予想数として計算した。図11に示すように、C p G 変異の予想数に対する確認された数の比率がメチル化レベルとともに上昇し、高メチル化レベルと低メチル化レベルとの間の確認された C p G T i 変異の数／予想された C p G T i 変異の数の比率には約5倍の変化があることが確認された。

【0091】

C p G サイトは2つのタイプ、すなわち、（1）高メチル化（平均化された場合、メチル化レベル>0.5）及び（2）低メチル化（平均化された場合、メチル化レベル≤0.5）に分類される。8つの C p G T i トリヌクレオチドコンテキストの各々についてのデノボ変異率を、高メチル化レベル及び低メチル化レベルのそれぞれに関して計算した。表

6に示すように、8つのCpGTiトリヌクレオチドコンテキスト全体で平均化すると、CpGTi変異率は高メチル化で $1.01e-07$ 、低メチル化で $2.264e-08$ であった。

【0092】

次いで、エクソーム配列決定データの対立遺伝子頻度スペクトル(AFS)を適合させた。あるそのようなサンプル実装形態では、100,000個の独立したサイトを仮定し、パラメータT1、r、及びnの様々な組み合わせを使用してシミュレーションが実行された。ここで、 $T1 \in (330, 350, 370, 400, 430, 450, 470, 500, 530, 550)$ 、 $r \in (20, 25, 30, \dots, 100, 105, 110)$ 、及び $n \in (1.0, 2.0, 3.0, 4.0, 5.0)$ を考慮した。

10

【0093】

異なるデノボ変異率、すなわちCpGTi、非CpGTi、及びTv(表6参照)を使用して、3つの主要クラスの各々を別々にシミュレートした。CpGTiについて、高いメチル化レベル及び低メチル化レベルを別々にシミュレートし、高メチル化サイト又は低メチル化サイトの割合を重みとして適用することによって2つのAFSを融合した。

【0094】

各パラメータ組み合わせ及び各変異率について、現在の日を通してヒト集団をシミュレートした。次いで、約246,000染色体の1,000セットをランダムにサンプリングした(ステップ288)(例えば、ターゲット世代又は最終世代290から)。これは、gnomADエクソームのサンプルサイズに対応する。1,000セットのそれぞれのサンプリングされたセット294にわたる平均化(ステップ292)によって、シミュレートされたAFS304を生成した。

20

【0095】

検証については、世界中の8つのサブ集団からの123,136人の個体の全エクソーム配列決定(whole-exome sequencing、WES)データを収集した(<http://gnomad.broadinstitute.org/>) genome Aggregation Database(gnomAD)v2.1.1からヒトエクソーム遺伝子多型データを取得した。フィルタを通過しなかった、カバレッジ中央値<15を有する、又は低複雑度領域若しくは部分重複領域(領域の境界は<https://storage.googleapis.com/gnomad-public/release/2.0.2/README.txt>で定義されている)内にある変異体は排除された。hg19ビルドのためのUCSCゲノムブラウザによって定義されるカノニカルコード配列にマッピングされた変異体は保持された。

30

【0096】

シングルトン、ダブルトン、 $3 \leq \text{アレルカウント}(AC) \leq 4$ 、 $\dots$ 、及び $33 \leq AC \leq 64$ を含む7つの対立遺伝子頻度カテゴリにおける同義変異体の数をカウントすることによって、gnomADの同義対立遺伝子頻度スペクトル308を生成した(ステップ306)。注目しているのはレア変異体であったため、 $AC > 64$ の変異体は破棄した。

【0097】

尤度比試験を適用して、3つの変異クラスにわたるレアな同義変異体のgnomAD AFS(すなわち、同義AFS308)へのシミュレートされたAFS304の適合を評価した(ステップ302)。図12A~図12Eに示されるように、ピアソンのカイ二乗統計量($-2 * \log$ 尤度比に対応)のヒートマップは、この例では最適なパラメータ組み合わせ(すなわち、最適なモデルパラメータ300)が $T1 = 530$ 、 $r = 30$ 、及び $n = 2.0$ で生じることを示している。図13は、このパラメータの組み合わせを使用してシミュレートされたAFS304が、確認されたgnomAD AFS(すなわち、同義AFS308)を模倣することを示している。推定される $T1 = 530$ 世代は、農業の広範な採用を約12,000年前(すなわち、新石器時代の始まり)とする考古学と合致する。センサス集団サイズと有効集団サイズとの間の比率は予想よりも低く、ヒト集団の多様性が実際にはかなり高いことを示唆する。

40

50

【0098】

図14を参照して、例示的な一実装形態では、順方向時間シミュレーションのコンテキストにおける選択効果に取り組むために、人間の拡大の歴史の最も可能性が高い人口統計モデルを先のシミュレーション結果において検索した。これらのモデルに基づいて、{0、0.0001、0.0002、・・・、0.8、0.9}から選択される選択係数320の異なる値を使用して、選択をシミュレーションに組み込んだ。各世代286について、親集団から突然変異を継承してデノボ変異を適用した後、選択係数320に従って変異の小さな割合をランダムにパーズした。

【0099】

シミュレーションの精度を改善するために、192個のトリヌクレオチドコンテキストの各々について、8071個のトリオ（すなわち、親子トリオ）から導出されたトリヌクレオチドコンテキストに固有のデノボ変異率230を使用して、別々のシミュレーションを適用した（ステップ282）。選択係数320の各値及び各変異率280の下で、約20,000染色体の初期サイズのヒト集団が現在の日まで拡大するようにシミュレートされた。結果として得られた集団（すなわち、ターゲット又は最終世代290）から1000組のセット294をランダムにサンプリングした（ステップ288）。各セットは約500,000個の染色体を含み、これはgnomAD+Topmed+UKバイオバンクのサンプルサイズに対応する。8つのCpGTiトリヌクレオチドコンテキストの各々について、高メチル化レベル及び低メチル化レベルを別々にシミュレートした。高メチル化サイト又は低メチル化サイトの割合を重みとして適用することによって2つのAFSを融合した。

【0100】

192個のトリヌクレオチドコンテキストのAFSが得られた後、エクソーム内の192個のトリヌクレオチドコンテキストの頻度によってこれらのAFSを重み付けして、エクソームのAFSを生成した。36個の選択係数の各々についてこの手順を繰り返した（ステップ306）。

【0101】

次いで、選択－枯渇曲線312を導出した。特に、変異に対する選択圧力が上昇するに伴い、変異体が徐々に枯渇すると予想される。中立進化（すなわち、選択無し）の下でのシナリオと比較して、浄化選択によって消去される変異体の割合を示すために、様々な選択レベル下でシミュレートされたAFS（304）を用いて、「枯渇」として特徴付けられるメトリックを定義した。この例では、枯渇は次のように特徴付けられ得る。

【0102】

【数1】

$$(1) \quad \text{枯渇} = 1 - \frac{\text{選択有りの変異体の数}}{\text{選択無しの変異体の数}}$$

【0103】

36個の選択係数の各々について枯渇値316を生成し（ステップ314）、図15に示される選択－枯渇曲線312を引いた。この曲線から、補間を適用し、任意の枯渇値に関連付けられた推定選択係数を取得することができる。

【0104】

順方向時間シミュレーションを使用した選択及び枯渇特性に関する上記議論に鑑み、これらのファクタを使用して、病原性（例えば、Primate AI又はpAI）スコアに基づきミスセンス変異体の選択係数を推定することができる。

【0105】

例として、図16を参照すると、122,439のgnomADエクソーム及び13,304のgnomAD全ゲノム配列決定（whole genome sequencing、WGS）サンプル（Topmedサンプルを除去した後）、約65KのTopmed WGSサンプ

ル、並びに約50KのUKバイオバンクWGSサンプルを含む約200,000人の個体から変異体対立遺伝子頻度を取得することによって、ある研究データが生成された（ステップ350）。各データセット内のレアな変異体（AF<0.1%）に焦点が置かれた。全ての変異体が、フィルタを通過すること、及びgnomADエクソームカバレッジに従いデプス中央値 ≥ 1 を有することが要求された。相乗係数 < -0.3 によって定義される、過剰のヘテロ接合体を示した変異体は除外した。全エクソーム配列決定に関して、ランダムフォレストモデルによって生成される確率が ≥ 0.1 である変異体は除外した。WGSサンプルに関して、ランダムフォレストモデルによって生成される確率が $0.4 \geq$ である変異体は除外した。タンパク質切断型変異体（protein-truncating variant、PTV）（ナンセンス変異を含む）に関して、スプライス変異体（すなわち、スプライシングドナー又はアクセプターサイトに生じる変異体）、及びフレームシフト、追加フィルタ、すなわち、機能喪失転写効果推定器（loss-of-function transcript effect estimator、LOFTEE）アルゴリズムによって推定された低信頼度に基づくフィルタリングが適用された。

10

【0106】

3つのデータセット間で変異体が融合され、枯渇メトリックを計算するために最終データセット352が形成された。以下の式に従い、3つのデータセット間で平均化するように対立遺伝子頻度が再計算された。

【0107】

【数2】

20

$$(2) \quad AF = \frac{\text{sum}(AC_i)}{\text{sum}(AN_i)}$$

式中、 i はデータセットインデックスを表す。あるデータセットに変異体が現れなかった場合、そのデータセットのためのACにはゼロ（0）が割り当てられた。

【0108】

ナンセンス変異及びスプライス変異体の枯渇に関して、各遺伝子における予想数と比較した、浄化選択によって枯渇したPTVの割合が計算され得る。遺伝子におけるフレームシフトの予想数を計算することの難しさのため、代わりに、機能喪失（loss-of-function、LOF）変異と称される、ナンセンス変異及びスプライス変異体に着目した。

30

【0109】

融合されたデータセットの各遺伝子においてナンセンス変異及びスプライス変異体の数をカウントし、確認されたLOF360の数を取得した（下式の分子）（ステップ356）。LOF364の予想数を計算するために（ステップ362）、制約メトリック（https://storage.googleapis.com/gnomad-public/release/2.1.1/constraint/gnomad.v2.1.1.lof_metrics.by_gene.txt.bgz）を含むファイルを、gnomADデータベースウェブサイトからダウンロードした。融合されたデータセットにおいて確認された同義変異体をベースラインとして使用し、LOFの予想数と、gnomADからの同義変異体の予想数との比率を掛けることによって、LOF364の予想数に変換した。次いで、枯渇メトリック380を計算し（ステップ378）、[0, 1]以内であることを検証した。0未満の場合には0が割り当てられ、逆もまた同様である。上記は、次のように表すことができる。

40

【0110】

【数3】

$$(3) \quad \text{枯渇}_{LOF} = 1 - \frac{\text{確認されたLOFの数}}{\text{予想されるLOFの数}}$$

ここで、

50

【0111】

【数4】

$$\text{予想されるLOFの数} = \text{確認された同義の数} * \frac{\text{gnomADの予想されるLOFの数}}{\text{gnomADの予想される同義の数}}$$

LOFの枯渇メトリック380に基づいて、各々の選択－枯渇曲線312を使用して、遺伝子ごとのLOF390の選択係数の推定値を導出することができる（ステップ388）。

【0112】

10

ミスセンス変異体における枯渇380の計算（ステップ378）に関して、（図17に示される）実装形態の一例では、遺伝子データセット418のために導出された予測病原性スコア（例えば、Primate AI又はpAISコア）のパーセンタイル420を使用して、各遺伝子内の全ての可能性のあるミスセンス変異体が表された。本明細書に記載される病原性スコアは変異体の相対的な適合度を測定するので、ミスセンス変異体の病原性スコアは、強い負の選択下の遺伝子においてより高くなる傾向があると予想され得る。逆に、適度な選択を有する遺伝子ではスコアが低くなると予想され得る。したがって、遺伝子に対する全体的な効果を回避するために、病原性スコア（例えば、pAISコア）のパーセンタイル420を使用することが適切である。

【0113】

20

各遺伝子について、病原性スコアパーセンタイル420（この例ではpAIパーセンタイル）を10本のビン（例えば、（0.0、0.1】、（0.1、0.2】、・・・、（0.9、1.0】）に分割し（ステップ424）、各ビンに含まれる確認されたミスセンス変異体の数428をカウントした（ステップ426）。枯渇メトリック380は、10本のビンの各々について枯渇メトリック380が計算されたことを除いて、LOFのものと同様に計算した430。本明細書に記載のLOF枯渇の計算に使用されるものと同様な、gnomADからのミスセンス／同義変異体の補正係数を、各ビンのミスセンス変異体の予想数に適用した。上記は、次のように表すことができる。

【0114】

【数5】

30

$$(4) \quad \text{枯渇}_{\text{ミスセンス}} = 1 - \frac{\text{各ビンにおける確認されたミスセンスの数}}{\text{予想されるミスセンスの数} / 10}$$

ここで、

【0115】

【数6】

$$\text{予想されるミスセンスの数} = \text{確認された同義の数} * \frac{\text{gnomADの予想されるミスセンスの数}}{\text{gnomADの予想される同義の数}}$$

【0116】

40

各遺伝子内の10本のビンの枯渇に基づいて、病原性スコアのパーセンタイル420と枯渇メトリック380との間の関係436が導出され得る（ステップ434）。一例では、各ビンのパーセンタイル中央値を求め、10本のビンの中央点に平滑スプラインを適合させた。この例が、それぞれ遺伝子の2つの例、具体的にはBRCA1及びLDLRに関して図18及び図19に示されており、両図は、枯渇メトリックが病原性スコアパーセンタイルに対して実質的に線形に増加することを示している。

【0117】

この方法論に基づいて、全ての可能なミスセンス変異体について、遺伝子固有の適合されたスプラインを使用して、各々の枯渇メトリック380を病原性パーセンタイルスコア420に基づいて予測することができる。選択－枯渇の関係（例えば、選択－枯渇曲線3

50

1 2 又は他の適合関数) を使用して、このミスセンス変異体の選択係数 3 2 0 を推定することができる。

【0 1 1 8】

更に、病的なレアミスセンス変異体及び P T V の予想数を個別に推定することができる。例えば、正常な個体のコードゲノムが有し得る病的なレア変異体の数の平均を推定することが望まれる可能性がある。この実装形態の例では、 $AF < 0.01\%$ のレア変異体に焦点を当てた。個体あたりの病的なレア P T V の予想数を計算することは、下式によって示されるように、特定の閾値を超える選択係数 3 2 0 を有する P T V の対立遺伝子頻度を足し合わせることに等価である。

【0 1 1 9】

【数 7】

$$(5) \quad E[\#PTV | s > k] = \sum AF | s > k$$

10

【0 1 2 0】

P T V はナンセンス変異、スプライス変異体、及びフレームシフトを含むため、カテゴリごとに計算を別々に行った。以下の表 6 に示される結果から、各個体が、 $s > 0.01$ (B R C A 1 変異と同じ又はより悪い) であるレア P T V を約 1.9 有することが確認され得る。

【0 1 2 1】

【表 6】

表 6. 異なる選択係数カットオフでの病的レア P T V の予想数

	$s > 0.2$	$s > 0.1$	$s > 0.05$	$s > 0.02$	$s > 0.01$	$s > 0.001$
E[#ナンセンス変異体]	0.0037	0.0209	0.0754	0.3078	0.6459	0.9959
E[#スプライス変異体]	0.0032	0.0153	0.0471	0.1639	0.3361	0.5100
E[#フレームシフト]	0.0234	0.0635	0.1574	0.4944	0.9338	1.3745
E[#PTV]	0.0303	0.0996	0.2799	0.9662	1.9159	2.8804

30

【0 1 2 2】

更に、異なる閾値を超える選択係数 3 2 0 を用いて、レアミスセンス対立遺伝子頻度を足し合わせることによって、病的レアミスセンス変異体の予想数を計算した。

【0 1 2 3】

【数 8】

$$(6) \quad E[\#ミスセンス | s > k] = \sum AF | s > k$$

【0 1 2 4】

以下の表 7 に示される結果から、 $s > 0.01$ の P T V が、ミスセンス変異体の約 4 倍、存在する。

【0 1 2 5】

40

50

【表 7】

表 7. 異なる選択係数カットオフでの病的ミスセンス変異体の予想数

E[#ミスセンス変異体]	$s > 0.2$	$s > 0.1$	$s > 0.05$	$s > 0.02$	$s > 0.01$	$s > 0.001$
gnomAD+Topmed補正あり	0.0022	0.0178	0.208	2.846	10.107	31.990
gnomAD+Topmed+UK バイオバンク 補正あり	0.00067	0.010	0.146	2.336	8.989	30.868
gnomAD+Topmed+UK バイオバンク 補正あり	0.0002	0.0035	0.0806	1.8327	8.2948	33.0164

10

【0126】

VI. 病原性スコアを使用した遺伝的疾患有病率の推定

臨床環境におけるミスセンス変異体の病原性スコアの採用及び使用を促進するために、臨床的に関心を持たれる遺伝子のうちの、病原性スコアと臨床疾患有病率との間の関係を調査した。特に、病原性スコアに基づく様々なメトリックを使用して遺伝的疾患の有病率を推定するための方法論が開発された。遺伝的疾患の有病率を予測するための病原性スコアに基づく2つのそのような方法論の非限定的な例が本明細書に記載されている。

【0127】

本研究で使用されるデータに関する予備的なコンテキストとして、本明細書で参照されるDiscovEHRデータはRegeneron Genetics CenterとGeisinger Health Systemとの間の連携によるものであり、Geisinger's MyCode Community Health Initiativeの50,726人の成人参加者の電子カルテ(electronic health record、EHR)からの全エクソーム配列決定を、臨床表現型と統合することによって、プレジジョンメディシンを触媒することを目的とする。これに鑑み、図20を参照すると、ACMG(American College of Medical Genetics and Genomics)の、臨床的にアクション可能な遺伝的所見の特定及び報告の推奨において特定されている56の遺伝子及び25の医学的状態を含む76の遺伝子(G76)が定義された(すなわち、遺伝子450のデータセット)。G76の遺伝子のうちの、ClinVar「病的」分類を有する変異体並びに既知の及び予測されるLOF変異体を含む、これらの潜在的に病的な変異体456の有病率を評価した。本明細書で論じられるように、各遺伝子におけるClinVar病原性変異体456の累積対立遺伝子頻度(CAF)466を導出した(ステップ460)。76の遺伝子のうちのほとんどについて、およその遺伝的疾患有病率が文献ソースから得られた。

20

30

【0128】

このコンテキストに鑑み、遺伝的疾患の有病率を予測するための病原性スコア206(例えば、PrimateAI又はpAISコア)に基づく手法の2つの例が開発された。図21を参照して、これらの方法論では、遺伝子固病原性スコア閾値212が用いられ(判定ブロック210)、遺伝子のミスセンス変異体200が病原性(すなわち、病原性変異体220)であるか否か(すなわち、非病原性変異体476)が決定される。一例では、予測病原性変異体について、病原性スコア206が特定の遺伝子における病原性スコアの75パーセンタイルよりも大きい場合に、カットオフが定義された。ただし、他のカットオフを適切に採用することができる。遺伝的疾患有病率メトリックは、ステップ478で導出されるように、予測病的ミスセンス変異体の予想累積対立遺伝子頻度(CAF)480として定義された。図22に示すように、Clinvar病原性変異体のDiscovEHR累積AFと、このメトリックとのスピアマン相関は0.5272である。同様に、図23は、疾患有病率とのこのメトリックのスピアマン相関が0.5954であることを示しており、良好な相関を示唆している。したがって、遺伝的疾患有病率メトリック

40

50

（すなわち、予測病的ミスセンス変異体の予想累積対立遺伝子頻度（CAF））は、遺伝的疾患有病率の予測因子として機能することができる。

【0129】

図21のステップ478として図示される、各遺伝子の遺伝的疾患有病率メトリックを計算することに関して、2つの異なる手法を評価した。第1の方法論では、図24を参照して、病的ミスセンス変異体220のリストのトリヌクレオチドコンテキスト構成500が最初を取得される（ステップ502）。本コンテキストにおいて、これは、可能なミスセンス変異体の全てを取得することに対応し得る。これらの病原性変異体220は、当該遺伝子の75パーセンタイル閾値（又は他の好適なカットオフ）を超える病原性スコア206を有するものである。

10

【0130】

各トリヌクレオチドコンテキスト500について、選択係数320が0.01であると仮定し、かつ当該トリヌクレオチドコンテキストのためにデノボ変異率280を使用して、本明細書に記載される順方向時間シミュレーションが実行され（ステップ502）、予想（すなわち、シミュレートされた）対立遺伝子頻度スペクトル（AFS）304が生成された。この方法論の一実装形態では、シミュレーションは、約400K染色体（約200Kサンプル）中の100,000個の独立したサイトをシミュレートした。したがって、そのようなコンテキストにおける特定のトリヌクレオチドコンテキスト500の予想AFS304は、シミュレートされたAFS/100,000*病原性変異体リスト内のトリヌクレオチドの発生である。192個のトリヌクレオチド全体を足し合わせることで、遺伝子の予想AFS304が生成される。この手法に係る特定の遺伝子（すなわち、予想CAF480）の遺伝的疾患有病率メトリックは、遺伝子の予想AFS304における、シミュレートされたレア対立遺伝子頻度（すなわち、 $AF \leq 0.001$ ）の和（ステップ506）として定義される。

20

【0131】

第2の方法論に従って導出される遺伝的疾患有病率メトリックは、第1の方法論を使用して導出されるものと同様であるが、病的ミスセンス変異体のリストの定義において異なる。第2の方法論によれば、図25を参照して、病原性変異体の予測枯渇が、当該遺伝子におけるタンパク質切断型変異体（PTV）の枯渇の75%以上である場合、病原性変異体220は遺伝子当たりの予測病原性変異体として定義される。例として、図25に示されるように、このコンテキストにおいて、関心のある変異体200の病原性スコア206が測定され得る（ステップ202）。病原性スコア206を使用して、本明細書で論じられるように、所定のパーセンタイルの病原性-枯渇の関係436を使用して、枯渇522を推定することができる（ステップ520）。次いで、推定枯渇522を枯渇閾値又はカットオフ524と比較して（判定ブロック526）、非病原性変異体476を病原性変異体220であると見なされるものから分けることができる。病原性変異体220が決定されると、処理は、ステップ478で上述したように、予想CAF480の導出に進み得る。

30

【0132】

この第2の方法論を使用して導出される遺伝的疾患有病率メトリックに関して、図26は、第2の方法論に従って計算された遺伝的疾患有病率のスピアマン相関を示し、Clinvar病原性変異体のDiscoverHR累積AFは0.5208である。同様に、図27は、第2の方法論に従って計算された遺伝的疾患有病率メトリックのスピアマン相関を示し、疾患有病率は0.4102であり、良好な相関を示す。したがって、第2の方法論を使用して計算されたメトリックも遺伝的疾患有病率の予測因子として機能し得る。

40

【0133】

VII. 病原性スコアの再校正

本明細書に記載されるように、本教示に従って生成される病原性スコアは、変異体の周りのDNAフランキング配列、複数の種の間での保全、及びタンパク質二次構造に主に基づいて訓練されたニューラルネットワークを使用して導出される。しかし、病原性スコア（

50

例えば、Primate AIスコア)に関連付けられたばらつきは大きい可能性がある(例えば、約0.15)。更に、病原性スコアを計算するために本明細書で論じられる一般化されたモデルの特定の実装形態は、訓練中にヒト集団において確認された対立遺伝子頻度の情報を利用しない。特定の状況において、高い病原性スコアを有するいくつかの変異体は、1を超える対立遺伝子数を有するようである可能性があり、これは、対立遺伝子数に基づいてこれらの病原性スコアにペナルティを課す必要があることを意味する。これに鑑み、そのような状況に対処するために病原性スコアを再校正することが有用であり得る。本明細書で論じられる1つの例示的な実施形態では、再校正手法は、変異体の病原性スコアのパーセンタイルに焦点を当ててもよい。なぜなら、これらはより堅牢で、遺伝子全体に加えられる選択圧力による影響を受けない可能性があるためである。

10

【0134】

これに鑑み、図28を参照して、再校正手法の一例では、確認された病原性スコアパーセンタイル550のノイズを評価及び考慮することを可能にするように、真の病原性パーセンタイルがモデル化される。このモデル化プロセスでは、真の病原性パーセンタイルは、離散的に均一に分布していると想定され得る(例えば、100個の値[0.01、0.02、...、0.99、1.00]をとる)。確認された病原性スコアパーセンタイル550は、標準偏差0.15の正規分布に続くいくつかのノイズ項を有する真の病原性スコアパーセンタイルを中心とすると仮定され得る。

(7) $obs\ AI \sim true\ AI + e, e \sim N(0, sd = 0.15)$

【0135】

20

このコンテキストにおける確認された病原性スコアパーセンタイル550の分布554は、図29に示されるように、ガウスノイズが重ねられた真の病原性スコアパーセンタイルの離散一様分布である。各線は、真の病原性スコアパーセンタイルの各値を中心とする正規分布を表す。図30は、ガウスノイズが重ねられた確認された病原性スコアパーセンタイルのこの離散一様分布556の密度プロットを示し、図31は、ステップ562で求められる累積分布関数(CDF)558を示す。このCDF558から、累積確率が100個の間隔に分割され、確認された病原性スコアパーセンタイル550のためにクォンタイル568が生成される(ステップ566)。

【0136】

真の病原性スコアパーセンタイル(図32のx軸)を有する変異体が、確認された病原性スコアパーセンタイル間隔(y軸)に含まれる確率を視覚化するために、この100×100確率行列の各列を正規化して足し合わせ、結果をヒートマップ572(図32)としてプロットする(ステップ570)ことができる。ヒートマップ572上の各点は、確認された病原性スコアパーセンタイル550間隔内の変異体が実際に真の病原性スコアパーセンタイルから来る確率(すなわち、真の病原性スコアパーセンタイル(x軸)を有する変異体が、確認された病原性スコアパーセンタイル間隔(y軸)に含まれる確率)を測定する。

30

【0137】

図33を参照すると、ミスセンス変異体に関して、本明細書に記載の方法論を使用して、各遺伝子における10本のピンのそれぞれについて枯渇メトリック522が求められた。この例では、本明細書の他の箇所で議論されているように、病原性スコア206は、ピニング処理の一部として関心のある変異体200のために計算され得る(ステップ202)。次に、それぞれの病原性スコア206を使用して、所定のパーセンタイルの病原性スコア-枯渇の関係436に基づき、枯渇522を推定することができる(ステップ520)。

40

【0138】

この枯渇メトリック522は、各ピンに含まれる変異体が浄化選択によって除去され得る確率を示す。例を遺伝子SCN2Aに関連して図34及び図35に示す。具体的には、図34は、SCN2A遺伝子のミスセンス変異体についての10本のピンのパーセンタイルにわたる枯渇確率を示す。変異体が選択を生き延びる確率は、生存確率580として示

50

され、ステップ 582 で求められる（1 - 枯渇）として定義され得る。この確率が 0.05 未満である場合、0.05 に設定されてもよい。図 35 は、SCN2A 遺伝子のミスセンス変異体についての 10 本のビンのパーセンタイルにわたる生存確率 580 を示す。両図において、x 軸上の 1.0 に記されている菱形は PTV を表す。

【0139】

一実装形態によれば、平滑スプラインが、生存確率対全てのビン（例えば、10 本のビン）にわたる各ビンの病原性スコアパーセンタイルの中央値に適合され、病原性スコアの各パーセンタイルの生存確率が生成された（ステップ 584）。この手法によれば、これは生存確率補正係数 590 を構成し、病原性スコア 206 のパーセンタイルが高いほど、変異体が浄化選択を生き延びる可能性が低いことを示唆する。他の実装形態では、平滑スプラインを適合させる代わりに補間などの他の技術が使用され得る。確認された変異体の病原性スコア 206 が高い場合、この補正係数 590 に従ってペナルティ化が課されるか、又は補正され得る。

10

【0140】

これに鑑み、図 36 を参照して、生存確率補正係数 590 を用いて再較正を行うことができる。例として、上記のような特定の遺伝子のヒートマップ 572 として視覚化され得る確率行列のコンテキストにおいて、ヒートマップ 572 の各行（例えば、寸法 50 × 50、100 × 100 などを含む確率行列）に、それぞれの生存確率補正係数 590（例えば、100 個の値のベクトル）を掛ける（ステップ 600）ことによって、その遺伝子の予想枯渇の分だけヒートマップ 572 の値を減らすことができる。次いで、足し合わせて 1 にするためにヒートマップの各行が再較正される。その後、図 37 に示すように、再較正されたヒートマップ 596 がプロット及び表示され得る。この例の再較正されたヒートマップ 596 は、x 軸上に真の病原性スコアパーセンタイルを表示し、再較正された確認された病原性スコアパーセンタイルを y 軸上に表示する。

20

【0141】

真の病原性スコアパーセンタイルは複数のビンに分割され（すなわち、1% ~ 10%（再較正ヒートマップされた 596 の最初の 10 本の列）が第 1 のビンに融合され、11% ~ 20%（再較正されたヒートマップ 596 の次の 10 本の列）が第 2 のビンに融合され、・・・など）（ステップ 604）、これは、ある変異体が各真の病原性スコアパーセンタイルビンから来る確率を表す。確認された病原性スコアパーセンタイル（例えば、再較正されたヒートマップ 596 の x 番目の行に対応する x%）を有する遺伝子の変異体について、この変異体が各真の病原性スコアパーセンタイルビン（例えば、本の 10 ビン）に含まれる確率を得ることができる（ステップ 608）。これは、各ビンへの変異体寄与 612 と称され得る。

30

【0142】

この例では、各ビン（例えば、10 本のビン）内のミスセンス変異体 620 の予想数（例えば、ステップ 624 で導出）は、それぞれの遺伝子において確認された全てのミスセンス変異体にわたる当該ビンへの変異体寄与の和である。遺伝子の各ビンに含まれるミスセンス変異体 620 のこの予想数に基づいて、本明細書で論じられるミスセンス変異体の枯渇式を使用して、各ミスセンスビンについて補正された枯渇メトリック 634 を計算することができる（ステップ 630）。これは図 38 に示されており、各パーセンタイルビンの補正された枯渇メトリックの例がプロットされている。具体的には、図 38 は、遺伝子 SCN2A における再較正された枯渇メトリックと元の枯渇メトリックとの比較を示す。x 軸上の 1.0 にプロットされた菱形は PTV の枯渇を示す。

40

【0143】

このように病原性スコア 206 を再較正すると、病原性スコア 206 のパーセンタイルにおけるノイズ分布がモデル化され、予測枯渇メトリック 522 におけるノイズが減少する。これは、本明細書で論じられるようなミスセンス変異体における選択係数 320 の推定に対するノイズの影響を軽減するのに役立ち得る。

【0144】

50

V I I I . ニューラルネットワークプライマー ニューラルネットワーク

上記において、ニューラルネットワークアーキテクチャ及び使用の様々な態様を、病原性分類又はスコアリングネットワークのコンテキストで言及してきた。本明細書で論じられるような病原性分類ネットワークを理解及び使用することを望む者にとって、これらのニューラルネットワークの設計及び使用の様々な態様の広範な知識は必要ないと思われるが、更なる詳細を望む者のために、追加の参照として以下のニューラルネットワークプライマーが提供される。

【 0 1 4 5 】

これに鑑み、一般的な意味での「ニューラルネットワーク」という用語は、対応する出力を受信するように訓練された計算構造物であって、その訓練に基づいて、入力修正、分類、又は他の方法で処理された、病原性スコアなどの出力を生成するものであると理解され得る。そのような構造物は、生物学的脳を模してモデル化されていることからニューラルネットワークと称され得る。構造物の様々なノードは「ニューロン」に相当し、ノード間の複雑な潜在的相互接続を可能にするために広範囲の他のノードと相互接続され得る。一般に、ニューラルネットワークは機械学習の一形態と見なすことができる。なぜなら、経路及び関連付けられたノードが例えば典型的には訓練され（例えば、入力及び出力が既知であるか、又はコスト関数が最適化され得るサンプルデータを使用して）、ニューラルネットワークが使用されて、その性能又は出力が修正又は再訓練されるにつれて経時的に学習又は進化することができるからである。

【 0 1 4 6 】

これに鑑み、更なる例示として、図 3 9 は、ニューラルネットワーク 7 0 0、ここでは、複数の層 7 0 2 を有する全結合ニューラルネットワーク 7 0 0 の例の簡略図を示す。本明細書に記載され、図 3 9 に示されるように、ニューラルネットワーク 7 0 0 は、相互接続された、メッセージを互いに交換する人工ニューロン 7 0 4（例えば a_1 、 a_2 、 a_3 ）のシステムである。例示されるニューラルネットワーク 7 0 0 は 3 つの入力と、隠れ層内の 2 つのニューロンと、出力層内の 2 つのニューロンとを有する。隠れ層は活性化関数 $f(\cdot)$ を有し、出力層は活性化関数 $g(\cdot)$ を有する。接続は、処理対象として訓練された入力を与えられたとき、適切に訓練されたネットワークが正しく応答するように訓練処理中に調整された、関連付けられた数値重みを有する（例えば、 w_{11} 、 w_{21} 、 w_{12} 、 w_{31} 、 w_{22} 、 w_{32} 、 v_{11} 、 v_{22} ）。入力層は生の入力を処理し、隠れ層は、入力層と隠れ層との間の接続の重みに基づいて、入力層からの出力を処理する。出力層は、隠れ層から出力を受け取り、隠れ層と出力層との間の接続の重みに基づいてそれを処理する。あるコンテキストでは、ネットワーク 7 0 0 は、特徴検出ニューロンの複数の層を含む。各層は、前の層からの異なる入力の組み合わせに応答する多数のニューロンを有する。これらの層は、第 1 の層が入力画像データ内のプリミティブパターンのセットを検出し、第 2 の層がパターンのパターンを検出し、第 3 の層がそれらのパターンのパターンを検出し、などというように構築され得る。

【 0 1 4 7 】

畳み込みニューラルネットワーク

ニューラルネットワーク 7 0 0 は、各自の動作モードに基づいて異なるタイプに分類され得る。例として、畳み込みニューラルネットワークは、dense 層又は密結合層とは対照的に、1 つ以上の畳み込み層を採用又は組み込むタイプのニューラルネットワークである。具体的には、密結合層は、入力特徴空間においてグローバルパターンを学習する。逆に、畳み込み層はローカルパターンを学習する。例として、画像の場合、畳み込み層は、入力の小さな窓又はサブセット内に認められるパターンを学習し得る。ローカルパターン又は特徴に焦点を当てることで、畳み込みニューラルネットワークが 2 つの有用な特性、すなわち、（1）学習するパターンが並進不変であり、（2）パターンの空間階層を学習することができる、という特性を有することができる。

【 0 1 4 8 】

第1の特性に関して、データセットの一部又はサブセットにおける特定のパターンを学習した後、畳み込み層は、同じ又は異なるデータセットの他の部分におけるパターンを認識することができる。対照的に、密結合ネットワークは、パターンが他の場所に（例えば、新しい場所）に存在する場合、パターンを新たに学習する必要がある。この特性により、他のコンテキスト及び場所における識別のために一般化され得る表現を学習するために必要な訓練サンプルが少なくなるため、畳み込みニューラルネットワークのデータ効率が良くなる。

【0149】

第2の特性に関して、第1の畳み込み層は小さなローカルパターンを学習する一方、第2の畳み込み層は、第1の層の特徴から作られたより大きなパターンを学習する（以下同様）ことができる。これにより、畳み込みニューラルネットワークは、一層複雑で抽象的な視覚的概念を効率的に学習することができる。

【0150】

これに鑑み、畳み込みニューラルネットワークは、多数の異なる層702内に配置された人工ニューロン704の複数の層を、複数の層を依存関係にする活性化関数を用いて相互接続することによって、非常に非線形のマッピングを学習することができる。1つ以上のサブサンプリング層及び非線形層が差し込まれた1つ以上の畳み込み層が含まれ、その後には通常、1つ以上の全結合層が続く。畳み込みニューラルネットワークの各要素は、前の層内の特徴のセットからの入力を受け取る。同じ特徴マップ内のニューロンは同じ重みを有するため、畳み込みニューラルネットワークは同時に学習する。これらのローカルに共有される重みはネットワークの複雑さを低減し、多次元入力データがネットワークに入ると、畳み込みニューラルネットワークは、特徴抽出及び回帰又は分類処理におけるデータ再構成の複雑さを回避する。

【0151】

畳み込みは、2つの空間軸（高さ及び幅）並びに深さ軸（チャンネル軸とも呼ばれる）を有する特徴マップと呼ばれる3Dテンソル上について演算を行う。畳み込み演算は、入力された特徴マップからパッチを抽出し、全てのパッチに同じ変換を適用して出力特徴マップを生成する。この出力特徴マップは依然として3Dテンソルであり、幅及び高さを有する。出力深さは層のパラメータであり、深さ軸における異なるチャンネルはフィルタを表すため、深さは任意であり得る。フィルタは、入力データの特定の側面を符号化する。

【0152】

例えば、第1の畳み込み層が所与のサイズ（28，28，1）の特徴マップを取得し、サイズが（26，26，32）である特徴マップを出力する例では、第1の畳み込み層は入力に対して32個のフィルタを計算する。これらの32個の出力チャンネルはそれぞれ、 26×26 の値グリッドを含む。これは、入力に対するフィルタの応答マップであり、入力内の異なる位置における当該フィルタパターンの応答を示す。すなわち、このコンテキストでの特徴マップ手段という用語は、深さ軸の全ての寸法が特徴（又はフィルタ）であり、2Dテンソル出力 $[:, :, n]$ は、入力に対するこのフィルタの応答の2D空間マップである。

【0153】

これに鑑み、畳み込みは2つのキーパラメータ、すなわち、（1）入力から抽出されたパッチのサイズ、及び（2）出力された特徴マップの深さ（すなわち、畳み込みによって計算されるフィルタの数）によって定義される。典型的な実装形態では、これらは深さ32から開始し、深さ64に続き、深さ128又は256で終了するが、実装形態によってはこの進行とは異なる可能性がある。

【0154】

図40を参照すると、畳み込み処理の視覚的概要が示されている。この例に示されるように、畳み込みは、3D入力特徴マップ720上でこれらの窓（例えば、サイズが 3×3 又は 5×5 の窓）をスライド（例えば、漸進的に移動）させることによって機能する。窓は全ての位置で停止し、周囲の特徴の3Dパッチ722（形状（window__height

10

20

30

40

50

height、width、input_depth))を抽出する。そのような3Dパッチ722はその後、(畳み込みカーネルと呼ばれる同一の学習された重み行列とのテンソル積を介して)形状(output_depth)の1Dベクトル724(すなわち、変換されたパッチ)に変換される。次いで、これらのベクトル724は、形状(height、width、output_depth)の3D出力特徴マップ726に空間的に再組み立てされる。出力特徴マップ726内の全ての空間位置が、入力特徴マップ720内の同じ位置に対応する。例えば、3×3窓の場合、ベクトル出力[i, j, :]は、3Dパッチ入力[i-1:i+1, j-1:j+1, :]に由来する。

【0155】

これに鑑み、畳み込みニューラルネットワークは、入力値と、訓練処理中に複数回の勾配更新反復にわたって学習される畳み込みフィルタ(重み行列)との間の畳み込み演算を実行する畳み込み層を含む。ここで、(m, n)はフィルタサイズであり、Wは重みの行列であり、畳み込み層は、ドット積 $W \cdot x + b$ を計算することによって、W及び入力Xの畳み込みを実行する。ここで、xはXのインスタンスであり、bはバイアスである。畳み込みフィルタが入力にわたってスライドする際のステップサイズはストライドと呼ばれ、フィルタ面積(m×n)は受容野と呼ばれる。入力の異なる位置にわたって同じ畳み込みフィルタが適用され、これは学習された重みの数を減らす。また、位置不変学習が可能になり、すなわち、入力内に重要なパターンが存在する場合、畳み込みフィルタは、それが配列内のどこにあるかとそれを学習する。

【0156】

畳み込みニューラルネットワークの訓練

上記から理解され得るように、畳み込みニューラルネットワークの訓練は、関心のある所与のタスクを実行するネットワークの重要な態様である。畳み込みニューラルネットワークは、入力データから特定の出力推定結果が得られるように調整又は訓練される。畳み込みニューラルネットワークは、出力推定結果が正解(ground truth)に徐々に一致する又は近づくまで、出力推定結果及び正解の比較に基づく逆伝搬を使用して調整される。

【0157】

畳み込みニューラルネットワークは、正解と実際の出力との間の差(すなわち、誤差δ)に基づき、ニューロン間の重みを調整することによって訓練される。訓練処理における中間ステップは、本明細書に記載されるように、畳み込み層を使用して入力データから特徴ベクトルを生成することを含む。出力から開始し、各層の重みに対する勾配が計算される。これはバックワードパス又は後方に進む、と表現される。ネットワーク内の重みは、負の勾配と前の重みとの組み合わせを使用して更新される。

【0158】

一実装形態では、畳み込みニューラルネットワーク150は、勾配降下による誤差の逆伝搬を実行する確率的勾配更新アルゴリズム(例えば、ADAM)を使用する。アルゴリズムは、ネットワーク内の全てのニューロンの活性化を計算し、フォワードパスのための出力を生成することを含む。その後、層ごとに誤差及び正しい重みが計算される。一実装形態では、畳み込みニューラルネットワークは、勾配降下最適化を使用して、全ての層にわたる誤差を計算する。

【0159】

一実装形態では、畳み込みニューラルネットワークは、確率的勾配降下(stochastic gradient descent、SGD)を使用してコスト関数を計算する。SGDは、たった1つのランダムデータペアからそれを計算することによって、損失関数内の重みに関する勾配に近似する。他の実装形態では、畳み込みニューラルネットワークは、ユークリッド損失及びソフトマックス損失などの異なる損失関数を使用する。更なる実施形態では、畳み込みニューラルネットワークはAdam確率的オプティマイザを使用する。

【0160】

畳み込み層

畳み込みニューラルネットワークの畳み込み層は特徴抽出器として機能する。具体的には、畳み込み層は、入力データを学習して階層特徴に分解することができる適合型特徴量抽出器として機能する。畳み込み演算は典型的には「カーネル」を含み、これは入力データにフィルタとして適用され、出力データが生成される。

【0161】

畳み込み演算は、入力データに対してカーネルをスライド（例えば、漸進的に移動）させることを含む。カーネルの各位置について、カーネル及び入力データの重なっている値が乗算され、結果が加算される。積の和は、カーネルが中心とする入力データの点における出力データの値である。多数のカーネルから得られた異なる出力は特徴マップと呼ばれる。

10

【0162】

畳み込み層が訓練されると、新しい推測データに対して認識タスクを実行するために適用される。畳み込み層は訓練データから学習されるので、明示的な特徴抽出が回避され、訓練データから暗示的に学習される。畳み込み層は、訓練処理の一部として求められ、更新される畳み込みフィルタカーネル重みを使用する。畳み込み層は、入力の異なる特徴を抽出し、これらはより高い層で組み合わせられる。畳み込みニューラルネットワークは様々な数の畳み込み層を使用する。各層が、カーネルサイズ、パディング、特徴マップの数、及び重みなど、異なる畳み込みパラメータを有する。

【0163】

サブサンプリング層

20

畳み込みニューラルネットワークの実装形態の更なる態様は、層のサブサンプリングを含み得る。このコンテキストでは、サブサンプリング層は、畳み込み層によって抽出された特徴の解像度を下げることににより、抽出された特徴又は特徴マップをノイズ及び歪みに対してロバストにする。一実装形態では、サブサンプリング層は、平均プーリング及び最大値プーリングの二種類のプーリング演算を使用する。プーリング演算は、入力を重複していない空間又は領域に分割する。平均プーリングの場合、領域内の値の平均が計算される。最大値プーリングの場合、値の最大値が選択される。

【0164】

一実装形態では、サブサンプリング層は、前の層の出力を最大値プーリングにおける入力のうちの1つのみにマッピングすること、及び前の層の出力を平均プーリングにおける入力の平均にマッピングすることによって、前の層のニューロンのセットに対してプーリング演算を行うことを含む。最大値プーリングにおいて、プーリングニューロンの出力は入力内の最大値である。平均プーリングにおいて、プーリングニューロンの出力は、入力ニューロンセット内の入力値の平均値である。

30

【0165】

非線形層

本概念に関連するニューラルネットワーク実装形態の更なる態様は、非線形層の使用である。非線形層は、各隠れ層上の可能性の高い特徴の異なる識別を指し示すために、異なる非線形トリガ関数を使用する。非線形層は、非線形トリガを実装するために様々な特定の関数を使用し、限定はされないが、正規化線形関数（rectified linear unit、ReLU）、双曲線正接、絶対双曲線正接、シグモイド、及び連続トリガ（非線形）関数が挙げられる。一実装形態では、ReLU活性化が $y = \max(x, 0)$ を実装し、層の入力サイズ及び出力サイズを同じに保つ。ReLUを使用することの1つの潜在的な利点は、畳み込みニューラルネットワークをより速く何度も訓練できる可能性があることである。ReLUは非連続な不飽和活性化関数であり、入力値がゼロより大きい場合、さもなければゼロである場合に入力に対して線形である。

40

【0166】

他の実装形態では、畳み込みニューラルネットワークは、連続的不飽和関数であるべきユニット活性化関数を使用することができる。べき活性化関数は、 c が奇数である場合には x 及び y 非対称活性化をもたらし、 c が偶数である場合には y 対称活性化をもたらすこ

50

とができる。一部の実装形態では、関数は非正規化線形活性化をもたらす。

【0167】

更に他の実装形態では、畳み込みニューラルネットワークは、連続的飽和関数であるシグモイドユニット活性化関数を使用することができる。シグモイドユニット活性化関数は負の活性化をもたらさず、y 軸に関してのみ非対称である。

【0168】

残差接続

畳み込みニューラルネットワークの更なる特徴は、図41に示すように、特徴マップの追加を介して下流に事前情報を再注入する残差接続の使用である。この例に示されるように、残差接続730は、データ処理フローに沿った情報損失を防止するのに役立つ、過去の出力テンソルを後の出力テンソルに追加することによって、以前の表現をデータの下流の流れに再注入することを含む。残差接続730は、先行する層の出力を、後続の層に入力として利用可能にすることを含み、ショートカットを連続的なネットワーク内で効果的に作成するというものである。先行する出力は、後のアクティブ化に連結されるのではなく、後のアクティブ化と合計される。なお、この場合には、両方のアクティブ化は、同じサイズであると仮定する。それらが異なるサイズの場合、目標の形状への先行のアクティブ化を再成形するために、線形変換を使用することができる。残差接続は、あらゆる大規模な深層学習モデルに存在し得る2つの問題、(1)勾配の消失と、(2)表現上のボトルネックとに対処するためのものである。一般に、10個より多くの層を有する任意のモデルに残差接続730を追加することが有益である可能性が高い。

【0169】

残差学習及びスキップ接続

本技術及び手法に関連する畳み込みニューラルネットワークに存在する別の概念は、スキップ接続の使用である。残差学習の背後にある原理は、残差マッピングの方が元のマッピングよりも学習しやすいことである。残差ネットワークは、学習精度の劣化を軽減するために、複数の残差ユニットを積み重ねる。残留ブロックは、ディープニューラルネットワークにおける勾配の消失に対抗するための特別な添加スキップ接続を使用する。残差ブロックの開始時に、データフローは2つのストリームに分離される：(1)第1のストリームは、ブロックの変化しない入力を担い、(2)第2は、重み及び非線形性を適用する。ブロックの終わりに、2つのストリームは、要素ごとの合計を使用してマージされる。そのような構成の1つの利点は、勾配がネットワークを介してより容易に流れることを可能にすることである。

【0170】

そのような残差ネットワークの利点によれば、深層畳み込みニューラルネットワーク(CNN)を容易に訓練することができ、また、データ分類や物体検出などの精度を高めることができる。畳み込みフィードフォワードネットワークは、 l 番目の層の出力を、入力として、 $(l+1)$ 番目の層に結合する。残差ブロックは、識別関数により非線形変換を迂回するスキップ接続を追加する。残差ブロックの利点は、勾配が、後の層から前の層まで、識別関数を直接流れることができることである。

【0171】

バッチ正規化

本発明の病原性分類手法に適用可能であり得る畳み込みニューラルネットワークの実装形態に関連する追加の態様はバッチ正規化である。バッチ正規化は、データ標準化をネットワークアーキテクチャの一体的な部分にすることによって、深層ネットワーク訓練を加速させるための方法である。バッチ正規化は、平均及び分散が訓練中に経時的に変化しても、データを適合的に正規化することができ、訓練中に見られるデータのバッチごとの平均及び分散の指数移動平均を内部的に維持することによって機能する。バッチ正規化の1つの効果は、残差接続のように勾配伝搬を助けることであり、したがって深層ネットワークの使用を容易化する。

【0172】

したがって、バッチ正規化は、全結合又は畳み込み層と同様に、モデルアーキテクチャに挿入することができる更なる別の層として見ることができる。バッチ正規化層は、典型的には、畳み込み又は密結合層の後に使用され得るが、畳み込み又は密結合層の前に使用することもできる。

【0173】

バッチ正規化は、入力をフィードフォワードして、バックワードパスを介してパラメータ及び自身の入力に関して勾配を計算するための定義を提供する。実際には、バッチ正規化層は、典型的には、畳み込み又は全結合層の後、かつ出力が活性化関数に供給される前に挿入される。畳み込み層の場合、同じ特徴マップ（すなわち、活性化）の異なる位置にある異なる要素は、畳み込み特性に従うために同じ方法で正規化される。したがって、活性化ごとにではなく、ミニバッチ内の全ての活性化が全ての位置にわたって正規化される。

10

【0174】

1D畳み込み

本手法に適用可能であり得る畳み込みニューラルネットワークの実装形態で使用される更なる技術は、配列からローカル1Dパッチ又は部分配列を抽出するための1D畳み込みの使用に関する。1D畳み込み手法は、入力配列内の窓又はパッチから各出力ステップを取得する。1D畳み込み層は、配列内のローカルパターンを認識する。全てのパッチで同じ入力変換が実行されるため、入力配列内の特定の位置で学習されたパターンを後で異なる位置にて認識することができ、並進のために1D畳み込み層を並進不変にすることができる。例えば、サイズ5の畳み込み窓を使用した塩基配列の1D畳み込み層処理は、長さ5以下の塩基又は塩基配列を学習することができるはずであり、入力配列において任意のコンテキストにおける塩基モチーフを認識することができるはずである。したがって、塩基レベル1D畳み込みは、塩基形態について学習することができる。

20

【0175】

グローバル平均プーリング

本コンテキストにおいて有用であり得るか、又は利用され得る畳み込みニューラルネットワークの別の態様は、グローバル平均プーリングに関する。具体的には、グローバル平均プーリングは、スコアリングのための最後の層内の特徴の空間平均を取ることによって、分類のための全結合（FC）層を置き換えるために使用され得る。これにより訓練負荷が低減され、過剰適合問題が回避される。グローバル平均プーリングは、モデルの前に構造を適用し、事前定義された重みを用いた線形変換と同等である。グローバル平均プーリングはパラメータの数を減らし、全結合層を排除する。全結合層は、典型的には、最も多くのパラメータ及び接続を有する層であり、グローバル平均プーリングは、同様の結果を達成するためのはるかに低いコストの手法を提供する。グローバル平均プーリングの主なアイデアは、各最後の層の特徴マップから平均値をスコアリングの信頼係数として生成し、ソフトマックス層に直接供給することである。

30

【0176】

グローバル平均プーリングは、限定されないが、（1）グローバル平均プーリング層に余分なパラメータが存在せず、したがって、グローバル平均プーリング層での過剰適合が回避される、（2）グローバル平均プーリングの出力が特徴マップ全体の平均であるため、グローバル平均プーリングは空間並進に対してロバストである、（3）全結合層内のパラメータの数が非常に大きく、ネットワーク全体の全パラメータの通常50%以上を占めるため、グローバル平均プーリング層によってそれらを置き換えることでモデルのサイズを大きく小さくすることができ、これはグローバル平均プーリングをモデル圧縮において非常に有用にする、という点を含む利益を提供し得る。

40

【0177】

最後の層におけるより強い特徴はより高い平均値を有すると期待されるため、グローバル平均プーリングは理にかなっている。一部の实装形態では、グローバル平均プーリングは、分類スコアのプロキシとして使用することができる。グローバル平均プーリング下の

50

特徴マップは、信頼マップ、及び特徴マップとカテゴリとの間の力の対応関係として解釈することができる。グローバル平均プーリングは、最後の層の特徴が直接分類のために十分に抽象的である場合に特に効果的であり得る。しかしながら、パーツモデルのように、マルチレベルの特徴をグループに組み合わせる必要がある場合にはグローバル平均プーリングのみで不十分又は好適ではない可能性があり、グローバル平均プーリングの後に単純な全結合層又は他の分類器を追加することによってより好適に対処され得る。

【0178】

IX. コンピュータシステム

理解され得るように、本議論のニューラルネットワークの態様、並びに記載のニューラルネットワークによって出力される病原性分類器に対して実行される分析及び処理は、コンピュータシステム又はシステム上に実装され得る。これに鑑み、更なるコンテキストとして、図42は、現在開示されている技術がその中で動作可能な例示的なコンピューティング環境800を示す。病原性分類器160、二次構造サブネットワーク130、及び溶媒露出度サブネットワーク132を有する深層畳み込みニューラルネットワーク102は、1つ以上の訓練サーバ802上で訓練される（サーバの数は、処理されるべきデータの量又は計算負荷に基づいて決定され得る）。訓練サーバによってアクセス、生成、及び／又は利用され得るこの手法の他の態様には、限定はされないが、訓練処理で使用される訓練データセット810、本明細書で論じられるような良性データセット生成器812、及び本明細書で論じられるような半教師あり学習器814が含まれる。訓練サーバの動作とのインタラクト及び／又はその制御を可能にするために、管理インターフェース816が設けられ得る。図42に示すように、訓練されたモデルの出力は、限定はされないが、生産環境の動作及び／又は試験で使用するために生産サーバ804に提供され得る試験データのセット820を含み得る。

【0179】

生産環境に関して、図42に示すように、病原性分類器160、二次構造サブネットワーク130、及び溶媒露出度サブネットワーク132を有する訓練された深層畳み込みニューラルネットワーク102は、クライアントインターフェース826を介してリクエストを出すクライアントから入力配列（例えば、生産データ824）を受け取る1つ以上の生産サーバ804上に展開される。生産サーバ804の数は、ユーザの人数、処理されるデータの量、又はより一般的に計算負荷に基づいて決定され得る。生産サーバ804は、病原性分類器160、二次構造サブネットワーク130、及び溶媒露出度サブネットワーク132のうちの少なくとも1つを介して入力配列を処理して、クライアントインターフェース826を介してクライアントに送られる出力（すなわち、病原性スコア又はクラスを含み得る推測データ828）を生成する。推測データ828は、本明細書で論じられるように、病原性スコア又は分類器、選択係数、枯渇メトリック、補正係数又は再較正されたメトリック、ヒートマップ、並びに対立遺伝子頻度及び累積対立遺伝子頻度などを含み得るが、これらに限定されない。

【0180】

訓練サーバ802、生産サーバ804、管理インターフェース816、及び／又はクライアントインターフェース826を実行又はサポートするために用いられ得る実際のハードウェアアーキテクチャに関して、そのようなハードウェアは、1つ以上のコンピュータシステム（例えば、サーバやワークステーションなど）として物理的に具現化されてもよい。そのようなコンピュータシステム850に認められ得る構成要素の例が図43に示されているが、この例は、そのようなシステムの全ての実施形態に認められない構成要素を含み得るか、又はそのようなシステムに認められ得る全ての構成要素を示さない可能性があることを理解されたい。更に、本手法の実際の態様では、部分的又は完全に仮想サーバ環境に、又はクラウドプラットフォームの一部として実装され得る。しかしながら、そのようなコンテキストでも、様々な仮想サーバインスタンス化は、図43に関して説明したようにハードウェアプラットフォーム上に実装される、ただし、記載されている特定の機能的態様は、仮想サーバインスタンスのレベルで実装されてもよい。

【0181】

これに鑑み、図43は、開示される技術を実装するために使用され得るコンピュータシステム850の簡略ブロック図である。コンピュータシステム850は、典型的には、バスサブシステム858を介して多数の周辺デバイスと通信する、少なくとも1つのプロセッサ854（例えば、CPU）を含む。これらの周辺デバイスは、例えば、記憶デバイス866（例えば、RAM874及びROM878）及びファイル記憶サブシステム870、ユーザインターフェース入力デバイス882、ユーザインターフェース出力デバイス886及びネットワークインターフェースサブシステム890を含む記憶サブシステム862を含むことができる。入力デバイス及び出力デバイスは、コンピュータシステム850とのユーザ対話を可能にする。ネットワークインターフェースサブシステム890は、他のコンピュータシステム内の対応するインターフェースデバイスへのインターフェースを含む外部ネットワークへのインターフェースを提供する。

10

【0182】

コンピュータシステム850が病原性分類器を実装又は訓練するために使用される一実装形態では、良性データセット生成器812、変異体病原性分類器160、二次構造分類器130、溶媒露出度分類器132、及び半教師あり学習器814などのニューラルネットワーク102は、記憶サブシステム862及びユーザインターフェース入力デバイス882に通信可能にリンクされている。

【0183】

図示の例、及びコンピュータシステム850が本明細書で論じられるようにニューラルネットワークを実装又は訓練するために使用されるコンテキストにおいて、1つ以上の深層学習プロセッサ894が、コンピュータシステム850の一部として存在してもよく、又は他の方法でコンピュータシステム850と通信してもよい。そのような実施形態では、深層学習プロセッサは、GPU又はFPGAであり得、Google Cloud Platform、Xilinx、及びCirrasculeなどの深層学習クラウドプラットフォームによってホストされ得る。深層学習プロセッサの例は、GoogleのTensor Processing Unit (TPU)、GX4 Rackmount Series、GX8 Rackmount Seriesのようなラックマウントソリューション、NVIDIA DGX-1、Microsoft' Stratix V FPGA、GraphcoreのIntelligent Processor Unit (IPU)、Snapdragon processorsを有するQualcommのZeroth Platform、NVIDIAのVolta、NVIDIAのDRIVE PX、NVIDIAのJETSON TX1/TX2 MODULE、IntelのNirvana、Movidius VPU、Fujitsu DPI、ARMのDynamicIQ、IBM TrueNorth、及び他のものを含む。

20

30

【0184】

コンピュータシステム850のコンテキストにおいて、ユーザインターフェース入力デバイス882はキーボードを含むことができる。マウス、トラックボール、タッチパッド、又はグラフィックタブレットなどのポインティングデバイス、スキャナ、ディスプレイに組み込まれたタッチスクリーン、音声認識システム及びマイクロフォンなどのオーディオ入力デバイス、並びに他のタイプの入力デバイスを含んでもよい。一般に、用語「入力デバイス」の使用は、コンピュータシステム850に情報を入力するための全ての可能なタイプのデバイス及び方式を包含するものとして解釈され得る。

40

【0185】

ユーザインターフェース出力デバイス886は、ディスプレイサブシステム、プリンタ、ファックス装置、又はオーディオ出力デバイスなどの非視覚ディスプレイを含むことができる。ディスプレイサブシステムは、陰極線管 (Cathode Ray Tube、CRT)、液晶ディスプレイ (Liquid Crystal Display、LCD) などのフラットパネルデバイス、投影デバイス、又は可視画像を生成するための何らかの他の機構を含んでもよい。ディスプレイサブシステムはまた、音声出力デバイスなどの非視覚ディスプレイを提供するこ

50

とができる。一般に、用語「出力デバイス」の使用は、コンピュータシステム 850 からユーザ又は別のマシン若しくはコンピュータシステムに情報を出力するための、全ての可能なタイプのデバイス及び方式を包含するものとして解釈され得る。

【0186】

記憶サブシステム 862 は、本明細書に記載されるモジュール及び方法のうちのいくつか又は全ての機能を提供するプログラミング及びデータ構築物を記憶する。これらのソフトウェアモジュールは、一般に、プロセッサ 854 単独で、又は他のプロセッサ 854 との組み合わせで実行される。

【0187】

記憶サブシステム 862 で使用されるメモリ 866 は、プログラム実行中に命令及びデータを記憶するためのメインランダムアクセスメモリ (random access memory、RAM) 878 と、固定命令が記憶された読み取り専用メモリ (read only memory、ROM) 874 とを含む多数のメモリを含むことができる。ファイル記憶サブシステム 870 は、プログラム及びデータファイルのための永続的な記憶装置を提供することができ、ハードディスクドライブ、関連する取り外し可能な媒体、CD-ROM ドライブ、光学ドライブ、又は取り外し可能な媒体カートリッジを含むことができる。特定の実施態様の機能を実装するモジュールは、記憶サブシステム 862 内のファイル記憶サブシステム 870 によって、又はプロセッサ 854 によってアクセス可能な他のマシン内に記憶され得る。

10

【0188】

バスサブシステム 858 は、コンピュータシステム 850 の様々な構成要素及びサブシステムを、意図されるように互いに通信させるための機構を提供する。バスサブシステム 858 は、単一のバスとして概略的に示されているが、バスサブシステム 858 の代替実施態様は、複数のバスを使用することができる。

20

【0189】

コンピュータシステム 850 自体は、パーソナルコンピュータ、ポータブルコンピュータ、ワークステーション、コンピュータ端末、ネットワークコンピュータ、テレビ、メインフレーム、スタンドアロンサーバ、サーバファーム、緩くネットワーク化されたコンピュータの緩く分散したセット、又は任意の他のデータ処理システム若しくはユーザデバイスを含む様々なタイプのものであり得る。コンピュータ及びネットワークは絶え間なく変化する性質のものであるため、図 43 に図示されるコンピュータシステム 850 の説明は、開示される技術を例示する目的のための具体例としてのみ意図されている。コンピュータシステム 850 の多くの他の構成は、図 43 に示されるコンピュータシステム 850 よりも多く、又は少ない構成要素を有することができる。

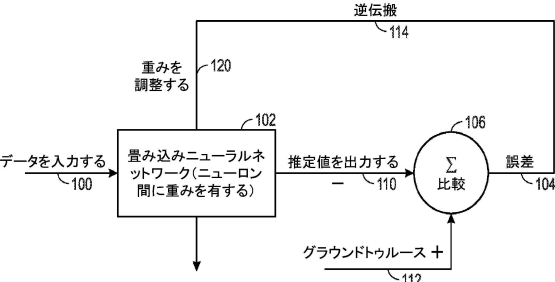
30

【0190】

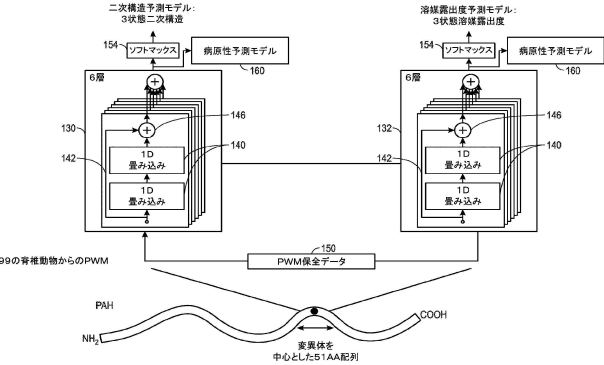
この書面による説明は、本発明を開示するために最良の態様を含む例を使用し、また、任意のデバイス又はシステムを作製及び使用し、任意の組み込まれた方法を実行することを含めて、あらゆる当業者が本発明を実践することを可能にする。本発明の特許性のある範囲は、特許請求の範囲によって定義され、当業者が想到する他の例を含み得る。そのような他の例は、これらが特許請求の範囲内の文字通りの言葉とは異ならない構造要素を含む場合、又はこれらが、特許請求の範囲内の文字通りの言葉とのごくわずかな違いを有する等価の構造要素を含む場合、特許請求の範囲内にあることを意図する。

40

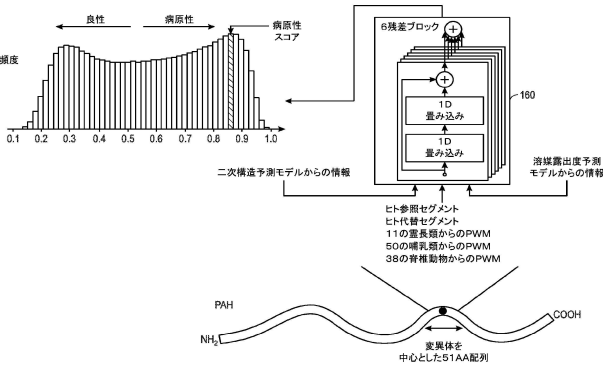
【図面】
【図 1】



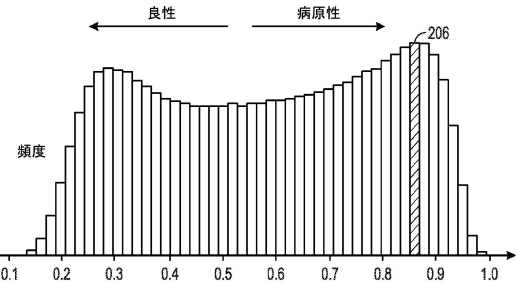
【図 2】



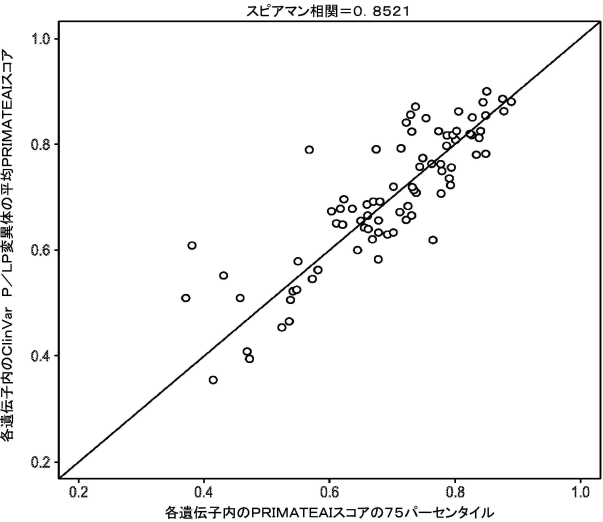
【図 3】



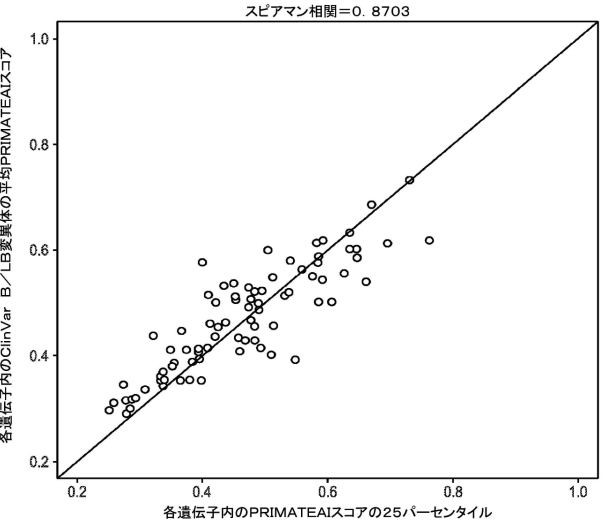
【図 4】



【図 5】



【図 6】



10

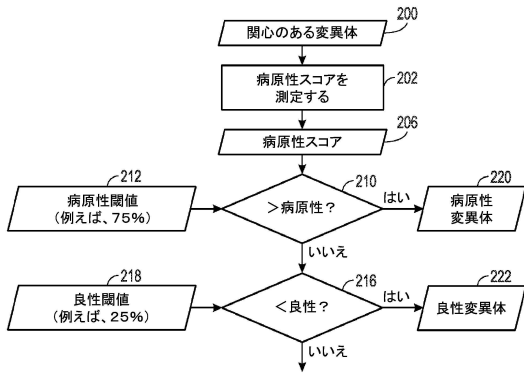
20

30

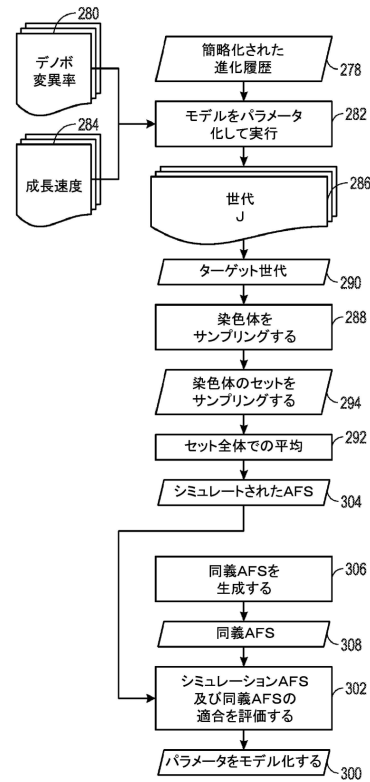
40

50

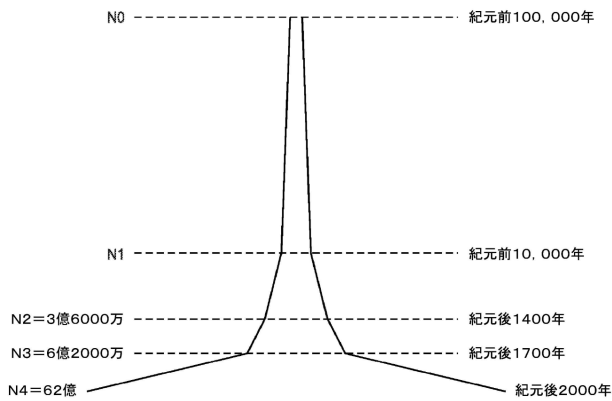
【図 7】



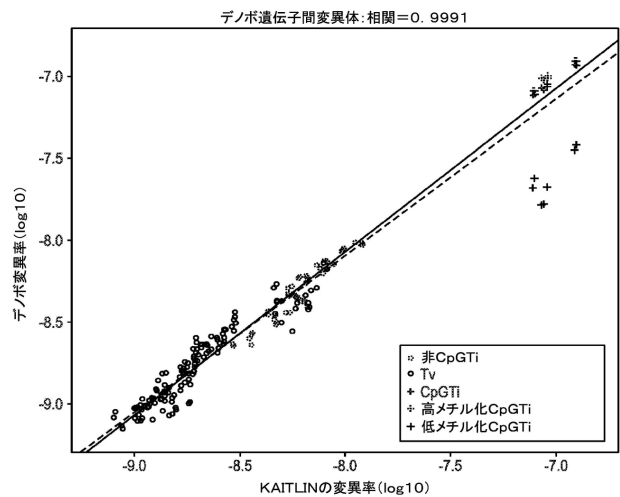
【図 8】



【図 9】



【図 10】



10

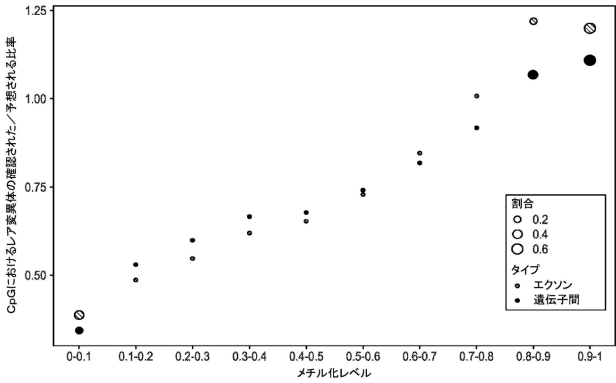
20

30

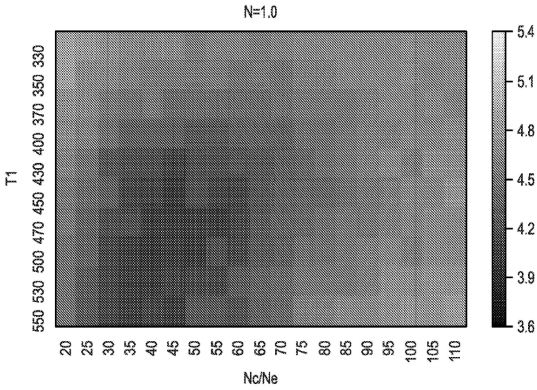
40

50

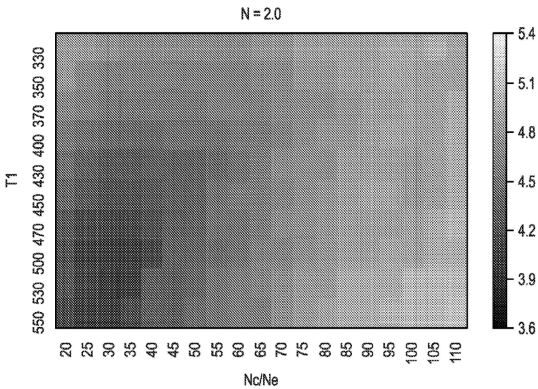
【図 1 1】



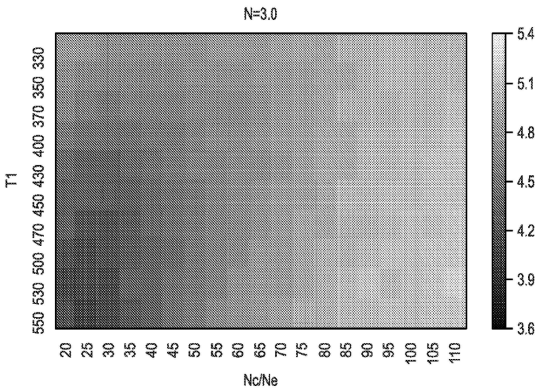
【図 1 2 A】



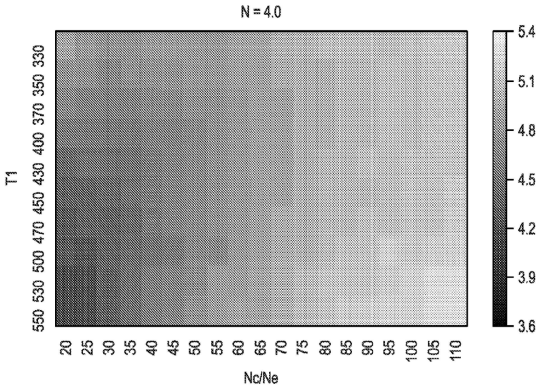
【図 1 2 B】



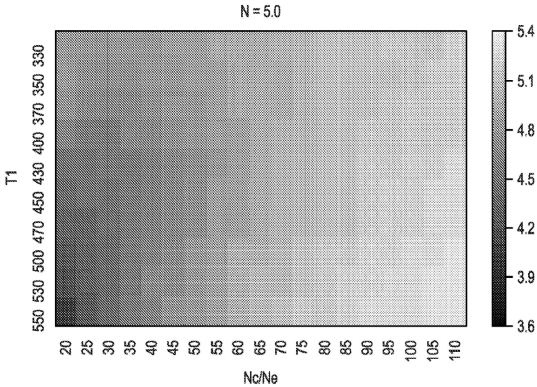
【図 1 2 C】



【図 1 2 D】



【図 1 2 E】



10

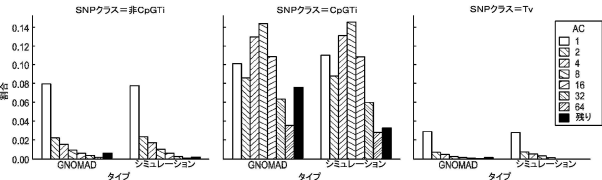
20

30

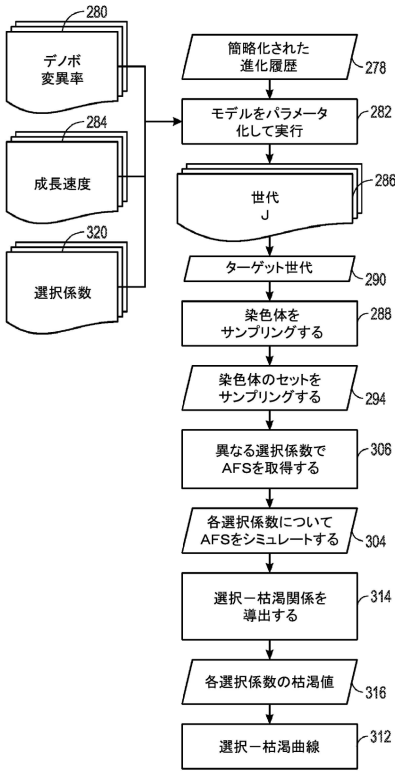
40

50

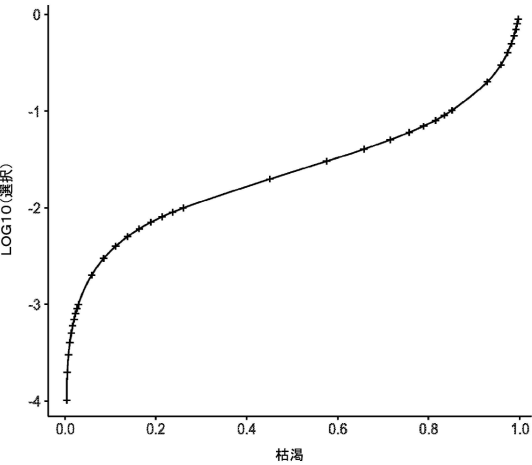
【図 13】



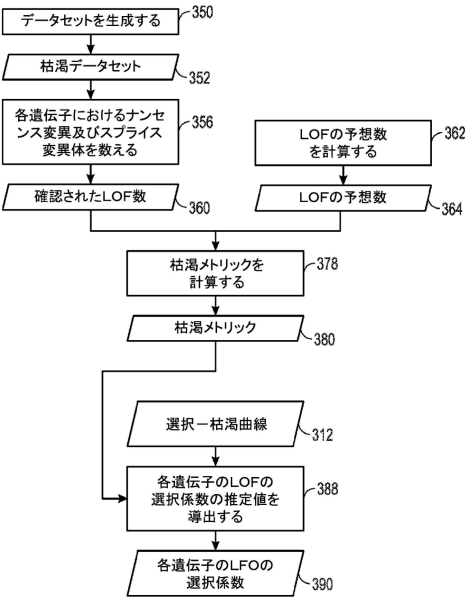
【図 14】



【図 15】



【図 16】



10

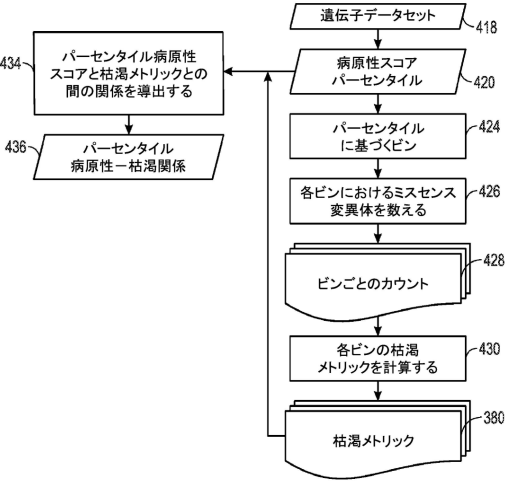
20

30

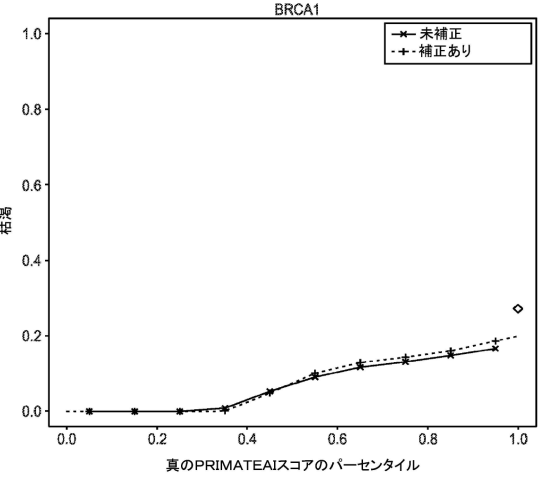
40

50

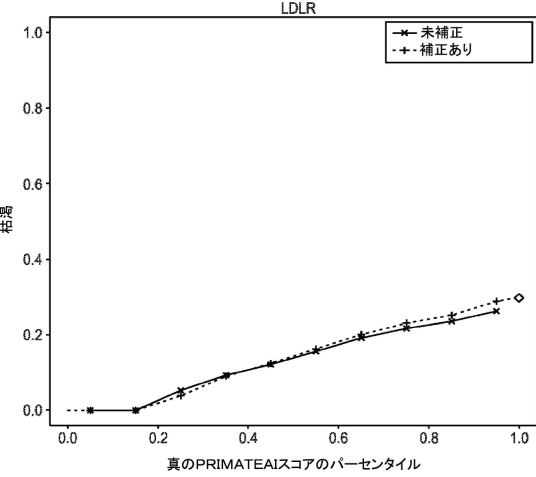
【図 17】



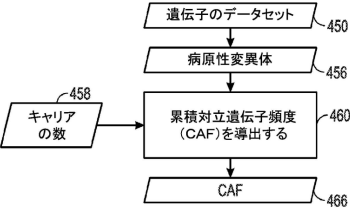
【図 18】



【図 19】



【図 20】



10

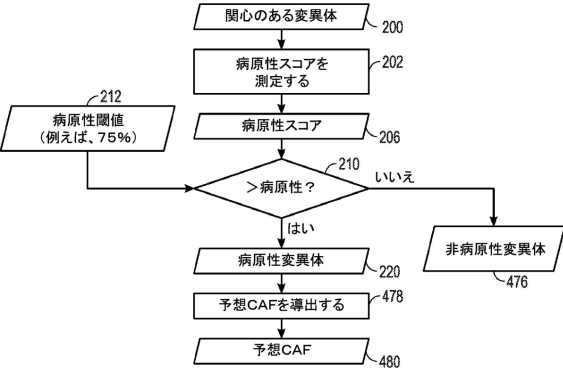
20

30

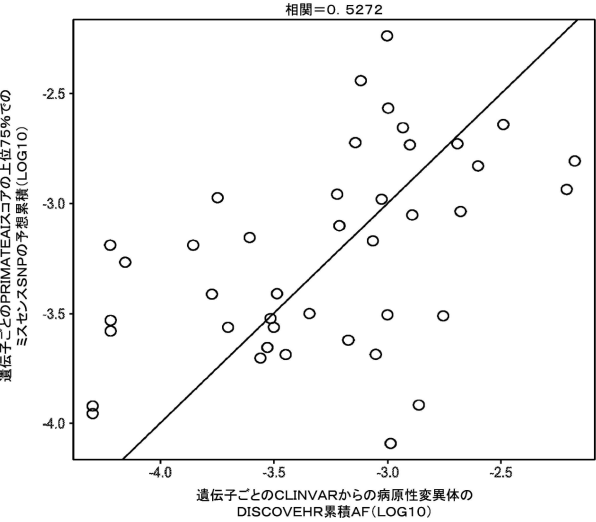
40

50

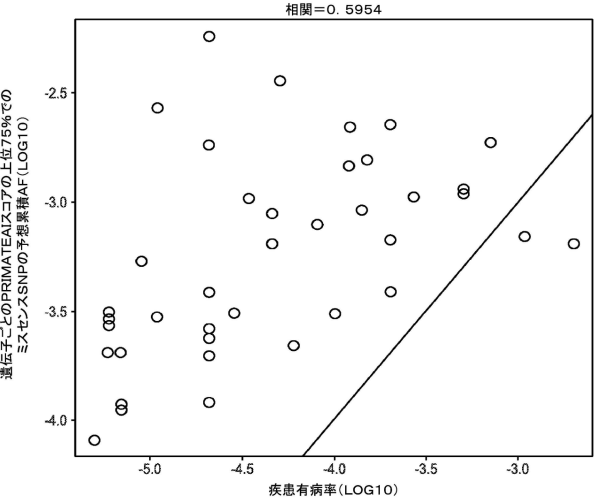
【図 2 1】



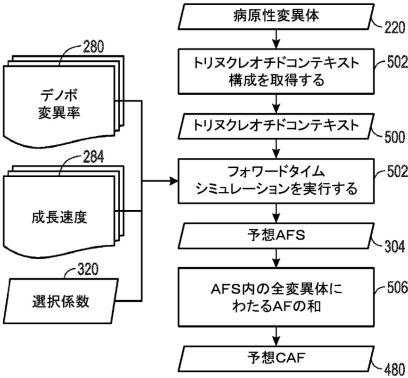
【図 2 2】



【図 2 3】



【図 2 4】



10

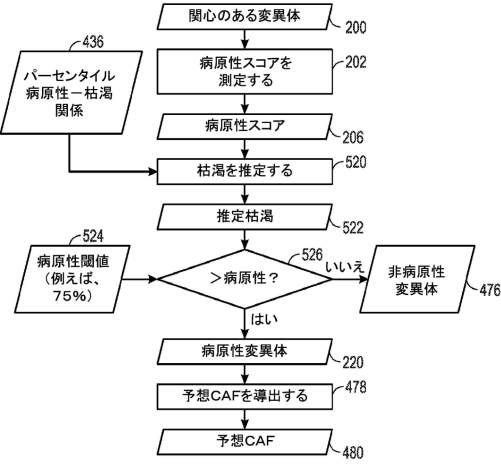
20

30

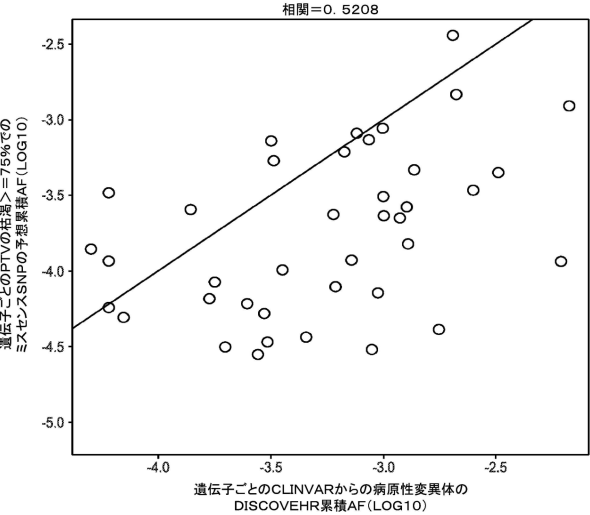
40

50

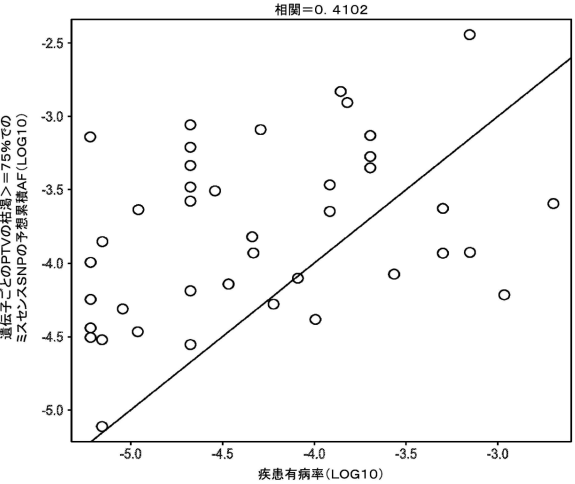
【図 25】



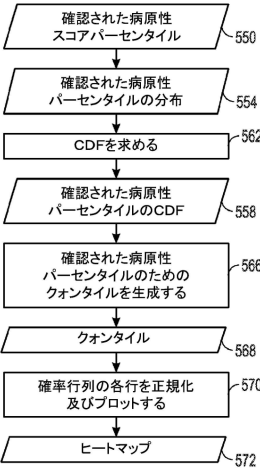
【図 26】



【図 27】



【図 28】



10

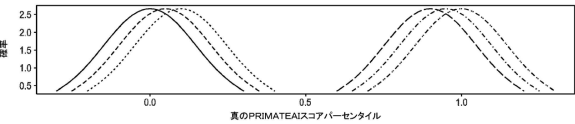
20

30

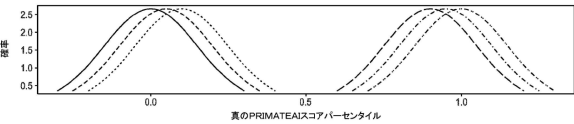
40

50

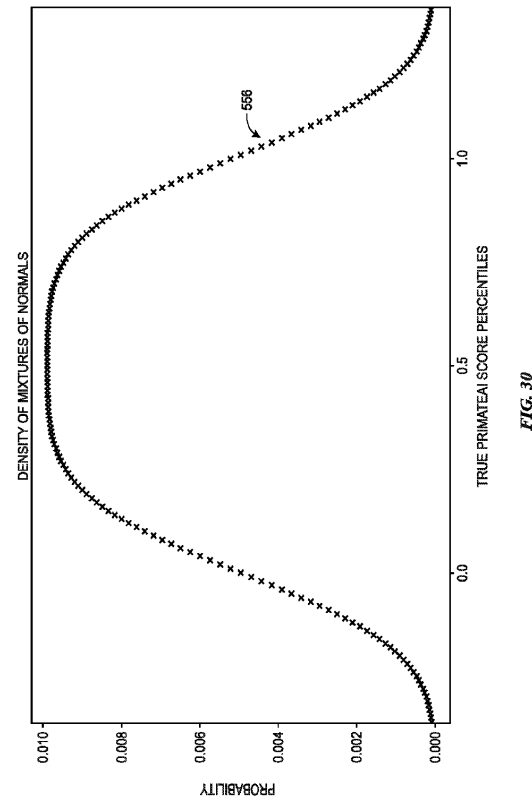
【図 29】



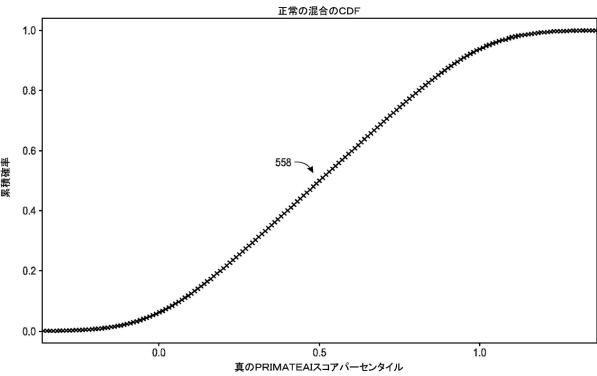
【図 30】



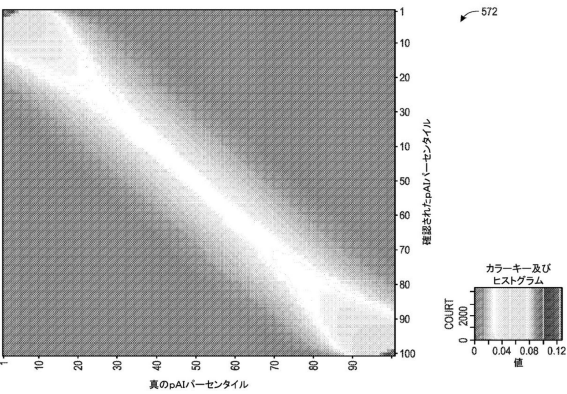
【図】



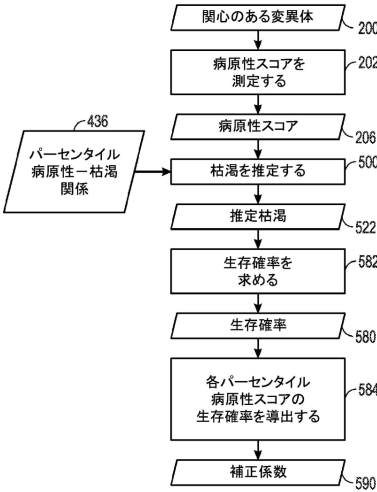
【図 31】



【図 32】



【図 33】



10

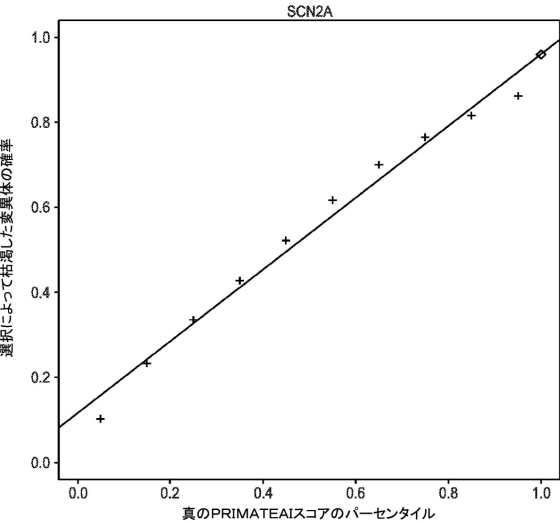
20

30

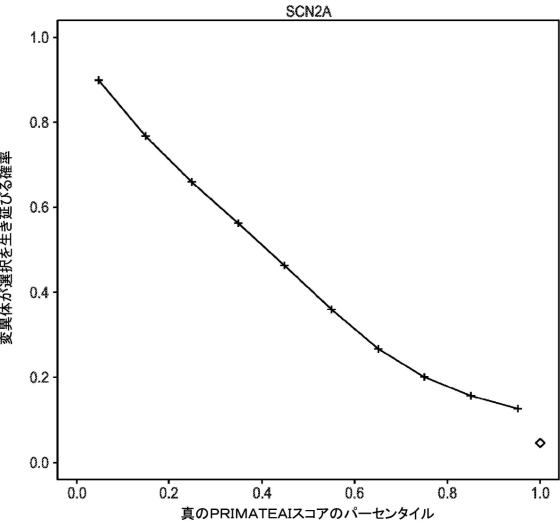
40

50

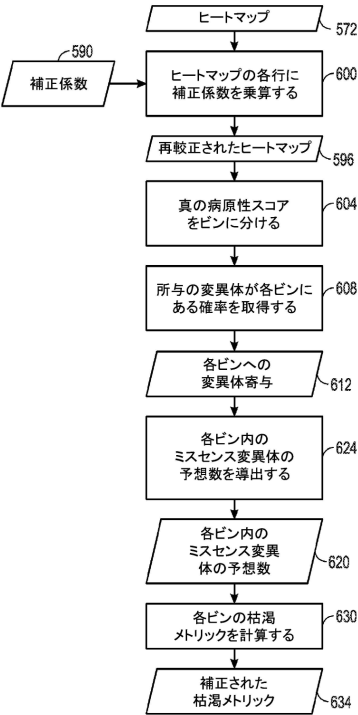
【図 3 4】



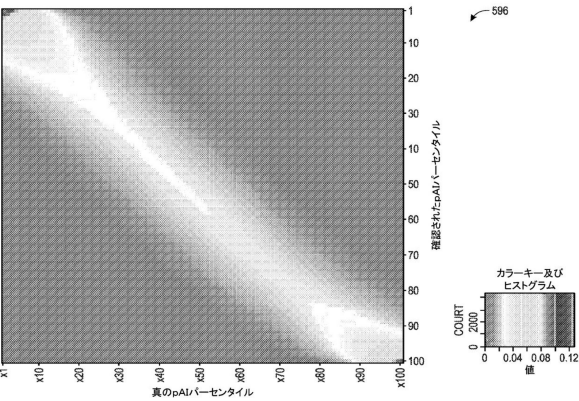
【図 3 5】



【図 3 6】



【図 3 7】



10

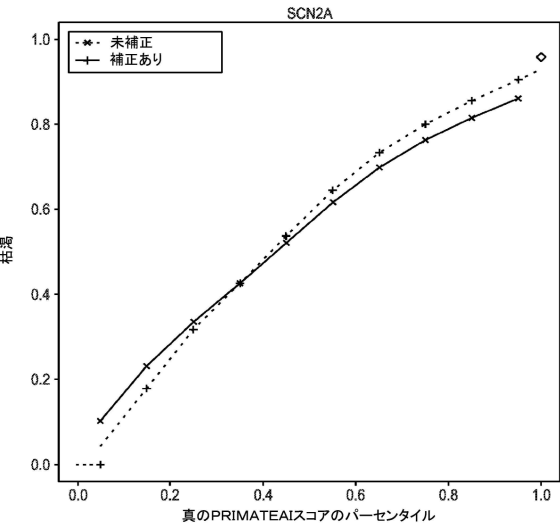
20

30

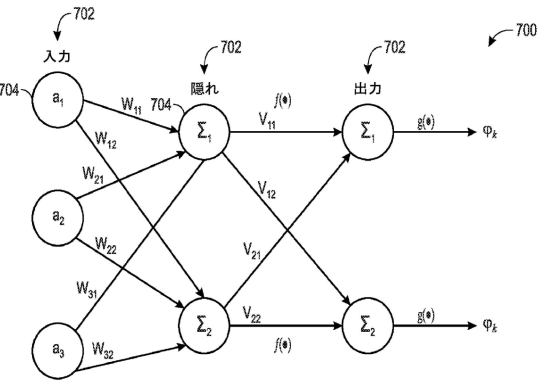
40

50

【図 3 8】

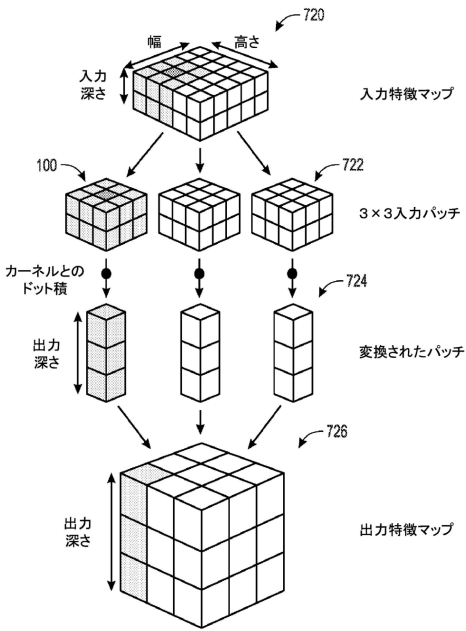


【図 3 9】

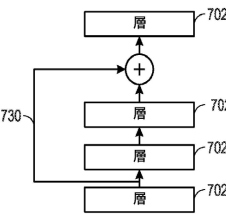


10

【図 4 0】



【図 4 1】



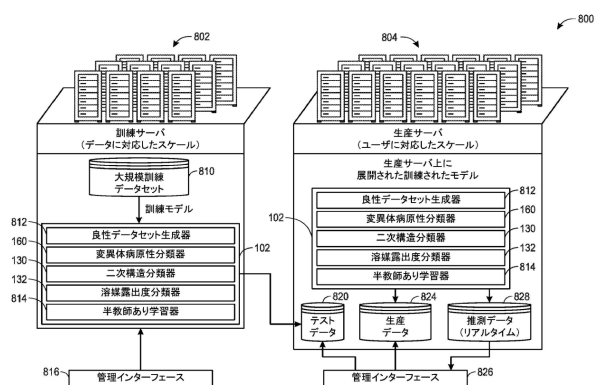
20

30

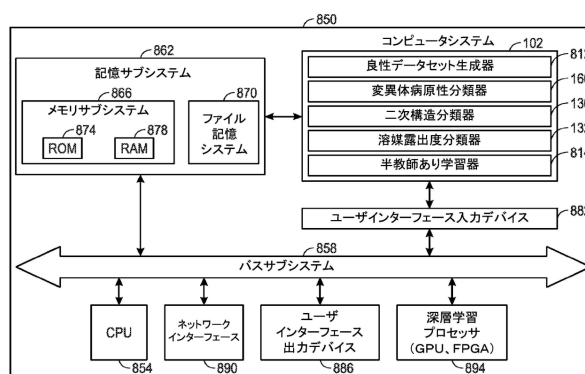
40

50

【圖 4 2】



【圖 4 3】



10

20

30

40

50

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2021/042599

A. CLASSIFICATION OF SUBJECT MATTER

INV. G16B20/20 G16B40/20

ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G16B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, BIOSIS, COMPENDEX, EMBASE, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	STEVEN FLYGARE ET AL: "The VAAST Variant Prioritizer (VVP): ultrafast, easy to use whole genome variant prioritization tool", BMC BIOINFORMATICS, BIOMED CENTRAL LTD, LONDON, UK, vol. 19, no. 1, 20 February 2018 (2018-02-20), pages 1-13, XP021253765, DOI: 10.1186/S12859-018-2056-Y	1,3,9, 10,13-21
Y	abstract page 2/13, column 1, paragraph 4 page 2/13, column 1, paragraph 6 - page 4/13, column 1, paragraph 2 page 6/13, column 1, paragraph 4 - page 7/13, column 2, paragraph 2; figure 3 page 9/13, column 2, paragraph 3 - page 10/13, column 1, paragraph 2 ----- -/--	2,4-8, 11,12

☒ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

15 October 2021

Date of mailing of the international search report

25/10/2021

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Thumb, Werner

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2021/042599

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	VAN DER VELDE K. JOERI ET AL: "GAVIN: Gene-Aware Variant Interpretation for medical sequencing", GENOME BIOLOGY, vol. 18, no. 1, 1 December 2017 (2017-12-01), XP55851226, DOI: 10.1186/s13059-016-1141-7 Retrieved from the Internet: URL:https://genomebiology.biomedcentral.com/track/pdf/10.1186/s13059-016-1141-7.pdf> abstract page 2/10, column 1, paragraph 3 - column 2, paragraph 1 page 3/10, column 1, paragraph 1 - page 4/10, column 1, paragraph 1 page 6/10, column 2, paragraph 1 - page 7/10, column 1, paragraph 1 -----	1,9,10
Y	CARON BARTHÉLÉMY ET AL: "NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans", GENOME BIOLOGY, vol. 20, no. 1, 1 December 2019 (2019-12-01), XP55851022, DOI: 10.1186/s13059-019-1634-2 Retrieved from the Internet: URL:https://genomebiology.biomedcentral.com/track/pdf/10.1186/s13059-019-1634-2.pdf> abstract page 4/22, column 1, paragraph 2 - page 9/22, column 1, paragraph 1 page 17/22, column 1, paragraph 2 - page 19/22, column 1, paragraph 1 -----	2
Y	US 2020/065675 A1 (SUNDARAM LAKSHMAN ET AL) 27 February 2020 (2020-02-27) paragraph [0040] paragraph [0212] - paragraph [0222] paragraph [0227] - paragraph [0228] paragraph [0251] - paragraph [0255] paragraph [0429] - paragraph [0453] -----	4-8,11,12

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2021/042599

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2020065675 A1	27-02-2020	AU 2018350891 A1	19-12-2019
		AU 2018352201 A1	19-12-2019
		AU 2018352203 A1	19-12-2019
		BR 112019027480 A2	19-05-2020
		CA 3066775 A1	25-04-2019
		CN 110800062 A	14-02-2020
		CN 110832596 A	21-02-2020
		CN 110870019 A	06-03-2020
		EP 3622520 A1	18-03-2020
		EP 3622521 A1	18-03-2020
		EP 3622523 A1	18-03-2020
		IL 271121 A	31-05-2021
		JP 6834029 B2	24-02-2021
		JP 2020525892 A	27-08-2020
		JP 2020530917 A	29-10-2020
		JP 2020530918 A	29-10-2020
		JP 2021093169 A	17-06-2021
		KR 20200006564 A	20-01-2020
		KR 20200011444 A	03-02-2020
		KR 20200011445 A	03-02-2020
		KR 20210000321 A	04-01-2021
		RU 2019139312 A	03-06-2021
		SG 11201912740T A	30-01-2020
		SG 11201912966U A	30-01-2020
		US 2019114511 A1	18-04-2019
		US 2019114544 A1	18-04-2019
		US 2019266491 A1	29-08-2019
		US 2020065675 A1	27-02-2020
		US 2020279157 A1	03-09-2020
		WO 2019079166 A1	25-04-2019
		WO 2019079180 A1	25-04-2019
		WO 2019079182 A1	25-04-2019

フロントページの続き

MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,N
E,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,
CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,IT,JO,JP,K
E,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,N
G,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,TM,
TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

(72)発明者 カイーハウ・ファー

アメリカ合衆国・カリフォルニア・9 2 1 2 2・サン・ディエゴ・イルミナ・ウェイ・5 2 0 0

(72)発明者 ジェレミー・フランシス・マクレ

アメリカ合衆国・カリフォルニア・9 2 1 2 2・サン・ディエゴ・イルミナ・ウェイ・5 2 0 0

F ターム (参考) 5L099 AA01