(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2020-525892
(P2020-525892A)
(43)公表日　令和2年8月27日(2020.8.27)

(51)Int.Cl.		F I		テーマコード（参考）
G16B 20/00	(2019.01)	G16B	20/00	
G06N 3/08	(2006.01)	G06N	3/08	

審査請求 有　予備審査請求 有　（全 157 頁）

(21)出願番号　特願2019-567719 (P2019-567719)
(86)(22)出願日　平成30年10月15日 (2018.10.15)
(85)翻訳文提出日　令和1年12月25日 (2019.12.25)
(86)国際出願番号　PCT/US2018/055840
(87)国際公開番号　WO2019/079166
(87)国際公開日　平成31年4月25日 (2019.4.25)
(31)優先権主張番号　62/573,144
(32)優先日　平成29年10月16日 (2017.10.16)
(33)優先権主張国・地域又は機関
米国 (US)
(31)優先権主張番号　62/573,149
(32)優先日　平成29年10月16日 (2017.10.16)
(33)優先権主張国・地域又は機関
米国 (US)

(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　９２１２２　サンディエゴ　イルミナ　ウェイ　５２００
(74)代理人　100108453
弁理士　村山　靖彦
(74)代理人　100110364
弁理士　実広　信哉
(74)代理人　100133400
弁理士　阿部　達彦
(72)発明者　ホン・ガオ
アメリカ合衆国・カリフォルニア・９２１２２・サン・ディエゴ・イルミナ・ウェイ・５２００

最終頁に続く

(54)【発明の名称】深層畳み込みニューラルネットワークを訓練するための深層学習ベースの技法

(57)【要約】
開示される技術は、バリアント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を構築することに関する。具体的には、開示される技術は、畳み込みニューラルネットワークベースの分類器の出力を対応するグラウンドトゥルースラベルと漸進的に照合する逆伝播ベースの勾配更新技法を使用して、訓練データについて畳み込みニューラルネットワークベースの分類器を訓練することに関する。畳み込みニューラルネットワークベースの分類器は、残差ブロックのグループを備え、残差ブロックの各グループは、残差ブロックの中の畳み込みフィルタの数、残差ブロックの畳み込みウィンドウサイズ、および残差ブロックの膨張畳み込み率によってパラメータ化され、畳み込みウィンドウのサイズは残差ブロックのグループ間で変動し、膨張畳み込み率は残差ブロックのグループ間で変動する。訓練データは、良性バリアントおよび病原性バリアントから生成された、翻訳された配列ペアの良性訓練例および病原性訓練例を含む。

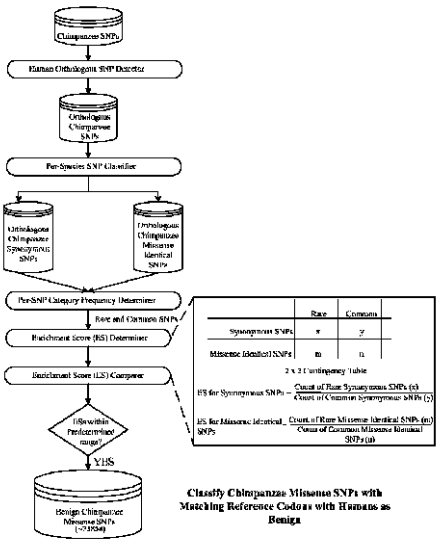


FIG. 46

【特許請求の範囲】
【請求項１】
バリアント病原性分類器を構築する方法であって、
基準タンパク質配列および代替タンパク質配列の良性訓練例のペアと病原性訓練例のペアとを入力として使用して、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バリアント病原性分類器を訓練するステップを含み、前記代替タンパク質配列が良性バリアントおよび病原性バリアントから生成され、
前記良性バリアントが、一般的なヒトミスセンスバリアントと、ヒトと一致する基準コドン配列を共有する代替的なヒト以外の霊長類コドン配列上で発生するヒト以外の霊長類ミスセンスバリアントとを含む、方法。
【請求項２】
前記一般的なヒトミスセンスバリアントが、少なくとも100000人のヒトからサンプリングされたヒト集団バリアントデータセットにわたって0.1%より高いマイナーアレル頻度(MAFと省略される)を有する、請求項1に記載の方法。
【請求項３】
前記サンプリングされたヒトが異なるヒト亜集団に属し、前記一般的なヒトミスセンスバリアントがそれぞれのヒト亜集団バリアントデータセット内で0.1%より高いMAFを有する、請求項2に記載の方法。
【請求項４】
前記ヒト亜集団が、アフリカ人/アフリカ系アメリカ人(AFRと省略される)、アメリカ人(AMRと省略される)、アシュケナージ系ユダヤ人(ASJと省略される)、東アジア人(EASと省略される)、フィンランド人(FINと省略される)、フィンランド人以外のヨーロッパ人(NFEと省略される)、南アジア人(SASと省略される)、および他(OTHと省略される)を含む、請求項3に記載の方法。
【請求項５】
前記ヒト以外の霊長類ミスセンスバリアントが、チンパンジー、ボノボ、ゴリラ、B.オランウータン、S.オランウータン、アカゲザル、およびマーモセットを含む、複数のヒト以外の霊長類の種からのミスセンスバリアントを含む、請求項1に記載の方法。
【請求項６】
エンリッチメント分析に基づいて、前記良性バリアントに特定のヒト以外の霊長類の種のミスセンスバリアントを含めるために、前記特定のヒト以外の霊長類の種を受け入れるステップをさらに含み、前記エンリッチメント分析が、前記特定のヒト以外の霊長類の種に対して、前記特定のヒト以外の霊長類の種の同義バリアントの第1のエンリッチメントスコアを前記特定のヒト以外の霊長類の種のミスセンス同一バリアントの第2のエンリッチメントスコアと比較することを含み、
ミスセンス同一バリアントが、ヒトと一致する基準コドン配列および代替コドン配列を共有するミスセンスバリアントであり、
前記第1のエンリッチメントスコアが、0.1%より高いMAFを伴う一般的な同義バリアントに対する0.1%より低いMAFを伴う稀な同義バリアントの比を決定することによって作り出され、
前記第2のエンリッチメントスコアが、0.1%より高いMAFを伴う一般的なミスセンス同一バリアントに対する0.1%より低いMAFを伴う稀なミスセンス同一バリアントの比を決定することによって作り出される、請求項1に記載の方法。
【請求項７】
稀な同義バリアントがシングルトンバリアントを含む、請求項6に記載の方法。
【請求項８】
前記第1のエンリッチメントスコアと前記第2のエンリッチメントスコアとの差が所定の範囲内にあり、前記良性バリアントに前記特定のヒト以外の霊長類のミスセンスバリアントを含めるために、前記特定のヒト以外の霊長類の種を受け入れるステップをさらに含む、請求項6に記載の方法。

【請求項９】
前記差が前記所定の範囲にあることが、前記ミスセンス同一バリアントが前記同義バリアントと同じ程度の自然選択を受けており、したがって前記同義バリアントと同じくらい良性であることを示す、請求項6に記載の方法。
【請求項１０】
前記エンリッチメント分析を繰り返し適用して、前記良性バリアントに前記ヒト以外の霊長類の種のミスセンスバリアントを含めるために複数のヒト以外の霊長類の種を受け入れるステップをさらに含む、請求項6に記載の方法。
【請求項１１】
前記ヒト以外の霊長類の種の各々に対する同義バリアントの第1のエンリッチメントスコアとミスセンス同一バリアントの第2のエンリッチメントスコアを比較するために、相同性のカイ二乗検定を使用するステップをさらに含む、請求項1に記載の方法。
【請求項１２】
前記ヒト以外の霊長類ミスセンスバリアントのカウントが少なくとも100000である、請求項1に記載の方法。
【請求項１３】
前記ヒト以外の霊長類ミスセンスバリアントの前記カウントが385236である、請求項12に記載の方法。
【請求項１４】
前記一般的なヒトミスセンスバリアントのカウントが少なくとも50000である、請求項1に記載の方法。
【請求項１５】
前記一般的なヒトミスセンスバリアントの前記カウントが83546である、請求項14に記載の方法。
【請求項１６】
バリアント病原性分類器を構築するためのコンピュータプログラム命令が焼かれた非一時的コンピュータ可読記憶媒体であって、プロセッサで実行されると、
基準タンパク質配列および代替タンパク質配列の良性訓練例のペアと病原性訓練例のペアとを入力として使用して、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バリアント病原性分類器を訓練するステップを含む方法を実施し、前記代替タンパク質配列が良性バリアントおよび病原性バリアントから生成され、
前記良性バリアントが、一般的なヒトミスセンスバリアントと、ヒトと一致する基準コドン配列を共有する代替的なヒト以外の霊長類コドン配列上で発生するヒト以外の霊長類ミスセンスバリアントとを含む、非一時的コンピュータ可読記憶媒体。
【請求項１７】
エンリッチメント分析に基づいて、前記良性バリアントに特定のヒト以外の霊長類の種のミスセンスバリアントを含めるために、前記特定のヒト以外の霊長類の種を受け入れることをさらに含む前記方法を実施し、前記エンリッチメント分析が、前記特定のヒト以外の霊長類の種に対して、前記特定のヒト以外の霊長類の種の同義バリアントの第1のエンリッチメントスコアを前記特定のヒト以外の霊長類の種のミスセンス同一バリアントの第2のエンリッチメントスコアと比較することを含み、
ミスセンス同一バリアントが、ヒトと一致する基準コドン配列および代替コドン配列を共有するミスセンスバリアントであり、
前記第1のエンリッチメントスコアが、0.1%より高いMAFを伴う一般的な同義バリアントに対する0.1%より低いMAFを伴う稀な同義バリアントの比を決定することによって作り出され、
前記第2のエンリッチメントスコアが、0.1%より高いMAFを伴う一般的なミスセンス同一バリアントに対する0.1%より低いMAFを伴う稀なミスセンス同一バリアントの比を決定することによって作り出される、請求項16に記載の非一時的コンピュータ可読記憶媒体。

【請求項１８】
前記ヒト以外の霊長類の種の各々に対する同義バリアントの第1のエンリッチメントスコアとミスセンス同一バリアントの第2のエンリッチメントスコアを比較するために、相同性のカイ二乗検定を使用することをさらに含む前記方法を実施する、請求項16に記載の非一時的コンピュータ可読記憶媒体。
【請求項１９】
バリアント病原性分類器を構築するためのコンピュータ命令がロードされたメモリに結合された1つまたは複数のプロセッサを含むシステムであって、前記プロセッサで実行されると、前記命令が、
基準タンパク質配列および代替タンパク質配列の良性訓練例のペアと病原性訓練例のペアとを入力として使用して、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バリアント病原性分類器を訓練することを含む活動を実施し、前記代替タンパク質配列が良性バリアントおよび病原性バリアントから生成され、
前記良性バリアントが、一般的なヒトミスセンスバリアントと、ヒトと一致する基準コドン配列を共有する代替的なヒト以外の霊長類コドン配列上で発生するヒト以外の霊長類ミスセンスバリアントとを含む、システム。
【請求項２０】
エンリッチメント分析に基づいて、前記良性バリアントに特定のヒト以外の霊長類の種のミスセンスバリアントを含めるために、前記特定のヒト以外の霊長類の種を受け入れる活動をさらに実施し、前記エンリッチメント分析が、前記特定のヒト以外の霊長類の種に対して、前記特定のヒト以外の霊長類の種の同義バリアントの第1のエンリッチメントスコアを前記特定のヒト以外の霊長類の種のミスセンス同一バリアントの第2のエンリッチメントスコアと比較することを含み、
ミスセンス同一バリアントが、ヒトと一致する基準コドン配列および代替コドン配列を共有するミスセンスバリアントであり、
前記第1のエンリッチメントスコアが、0.1%より高いMAFを伴う一般的な同義バリアントに対する0.1%より低いMAFを伴う稀な同義バリアントの比を決定することによって作り出され、
前記第2のエンリッチメントスコアが、0.1%より高いMAFを伴う一般的なミスセンス同一バリアントに対する0.1%より低いMAFを伴う稀なミスセンス同一バリアントの比を決定することによって作り出される、請求項19に記載のシステム。
【請求項２１】
バリアント分類のために畳み込みニューラルネットワークベースの分類器を構築するコンピュータで実施される方法であって、
メモリに結合される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分類器を、前記畳み込みニューラルネットワークベースの分類器の出力を対応するグラウンドトゥルースラベルと漸進的に照合する逆伝播ベースの勾配更新技法を使用して、訓練データについて訓練するステップを含み、
前記畳み込みニューラルネットワークベースの分類器が残差ブロックのグループを備え、
残差ブロックの各グループが、前記残差ブロックの中の畳み込みフィルタの数、前記残差ブロックの畳み込みウィンドウサイズ、および前記残差ブロックの膨張畳み込み率によってパラメータ化され、
前記畳み込みウィンドウサイズが残差ブロックの前記グループ間で変動し、
前記膨張畳み込み率が残差ブロックの前記グループ間で変動し、
前記訓練データが、良性バリアントおよび病原性バリアントから生成され良性訓練例および病原性訓練例として使用される、翻訳された配列のペアを含み、
前記良性バリアントが、一般的なヒトミスセンスバリアントと、ヒトと一致する基準塩基トリプレット配列を共有する代替的なヒト以外の霊長類塩基トリプレット配列上で発生

するヒト以外の霊長類ミスセンスバリアントとを含む、方法。
【請求項２２】
バリアント分類のために畳み込みニューラルネットワークベースの分類器を構築するためのコンピュータプログラム命令が焼かれた非一時的コンピュータ可読記憶媒体であって、プロセッサで実行されると、前記命令が、
メモリに結合される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分類器を、前記畳み込みニューラルネットワークベースの分類器の出力を対応するグラウンドトゥルースラベルと漸進的に照合する逆伝播ベースの勾配更新技法を使用して、訓練データについて訓練するステップを含む方法を実施し、
前記畳み込みニューラルネットワークベースの分類器が残差ブロックのグループを備え、
残差ブロックの各グループが、前記残差ブロックの中の畳み込みフィルタの数、前記残差ブロックの畳み込みウィンドウサイズ、および前記残差ブロックの膨張畳み込み率によってパラメータ化され、
前記畳み込みウィンドウサイズが残差ブロックの前記グループ間で変動し、
前記膨張畳み込み率が残差ブロックの前記グループ間で変動し、
前記訓練データが、良性バリアントおよび病原性バリアントから生成され良性訓練例および病原性訓練例として使用される、翻訳された配列のペアを含み、
前記良性バリアントが、一般的なヒトミスセンスバリアントと、ヒトと一致する基準塩基トリプレット配列を共有する代替的なヒト以外の霊長類塩基トリプレット配列上で発生するヒト以外の霊長類ミスセンスバリアントとを含む、非一時的コンピュータ可読記憶媒体。
【請求項２３】
バリアント分類のために畳み込みニューラルネットワークベースの分類器を構築するためのコンピュータ命令がロードされたメモリに結合される1つまたは複数のプロセッサを含むシステムであって、前記プロセッサで実行されると、前記命令が、
メモリに結合される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分類器を、前記畳み込みニューラルネットワークベースの分類器の出力を対応するグラウンドトゥルースラベルと漸進的に照合する逆伝播ベースの勾配更新技法を使用して、訓練データについて訓練することを含む活動を実施し、
前記畳み込みニューラルネットワークベースの分類器が残差ブロックのグループを備え、
残差ブロックの各グループが、前記残差ブロックの中の畳み込みフィルタの数、前記残差ブロックの畳み込みウィンドウサイズ、および前記残差ブロックの膨張畳み込み率によってパラメータ化され、
前記畳み込みウィンドウサイズが残差ブロックの前記グループ間で変動し、
前記膨張畳み込み率が残差ブロックの前記グループ間で変動し、
前記訓練データが、良性バリアントおよび病原性バリアントから生成され良性訓練例および病原性訓練例として使用される、翻訳された配列のペアを含み、
前記良性バリアントが、一般的なヒトミスセンスバリアントと、ヒトと一致する基準塩基トリプレット配列を共有する代替的なヒト以外の霊長類塩基トリプレット配列上で発生するヒト以外の霊長類ミスセンスバリアントとを含む、システム。
【発明の詳細な説明】
【技術分野】
【０００１】
付録
付録には、発明者らが著述した論文に列挙される潜在的な関連する参考文献の目録が含まれる。その論文の主題は、本出願がその優先権を主張する/その利益を主張する米国仮出願において扱われる。これらの参考文献は、要求に応じて訴訟代理人に対して利用可能にされることが可能であり、またはGlobal Dossierを介して入手可能であることがある。

その論文は最初の列挙される参考文献である。

【０００２】

優先出願

本出願は、2017年10月16日に出願された、Hong Gao、Kai-How Farh、Laksshman Sundaram、およびJeremy Francis McRaeによる「Training a Deep Pathogenicity Classifier Using Large-Scale Benign Training Data」という表題の米国仮特許出願第62/573,144号(代理人整理番号第ILLM 1000-1/IP-1611-PRV)、2017年10月16日に出願された、Kai-How Farh、Laksshman Sundaram、Samskruthi Reddy Padigepati、およびJeremy Francis McRaeによる「Pathogenicity Classifier Based On Deep Convolutional Neural Networks (CNNS)」という表題の米国仮特許出願第62/573,149号(代理人整理番号第ILLM 1000-2/IP-1612-PRV)、2017年10月16日に出願された、Hong Gao、Kai-How Farh、Laksshman Sundaram、およびJeremy Francis McRaeによる「Deep Semi-Supervised Learning that Generates Large-Scale Pathogenic Training Data」という表題の米国仮特許出願第62/573,153号(代理人整理番号第ILLM 1000-3 /IP-1613-PRV)、および、2017年11月7日に出願された、Hong Gao、Kai-How Farh、およびLaksshman Sundaramによる「Pathogenicity Classification of Genomic Data Using Deep Convolutional Neural Networks (CNNs)」という表題の米国仮特許出願第62/582,898号(代理人整理番号第ILLM 1000-4/IP-1618-PRV)の優先権または利益を主張する。これらの仮出願は、すべての目的のために本明細書において参照により引用される。

【０００３】

引用

以下は、本明細書に完全に記載されるかのようにすべての目的のために参照により引用される。

【０００４】

後にPCT出願公開第WO(未定)号として公開される、2018年10月15日に同時に出願された、Laksshman Sundaram、Kai-How Farh、Hong Gao、Samskruthi Reddy Padigepati、およびJeremy Francis McRaeによる、「DEEP CONVOLUTIONAL NEURAL NETWORKS FOR VARIANT CLASSIFICATION」という表題のPCT特許出願第PCT/US2018/(未定)号(代理人整理番号ILLM 1000-9/IP-1612-PCT)。

【０００５】

後にPCT出願第WO(未定)号として公開される、2018年10月15日に同時に出願された、Laksshman Sundaram、Kai-How Farh、Hong Gao、およびJeremy Francis McRaeによる、「SEMI-SUPERVISED LEARNING FOR TRAINING AN ENSEMBLE OF DEEP CONVOLUTIONAL NEURAL NETWORKS」という表題の国際特許出願第PCT/US18/(未定)号(代理人整理番号第ILLM 1000-10/IP-1613-PCT)。

【０００６】

同時に出願された、Hong Gao、Kai-How Farh、Laksshman Sundaram、およびJeremy Francis McRaeによる、「DEEP LEARNING-BASED TECHNIQUES FOR TRAINING DEEP CONVOLUTIONAL NEURAL NETWORKS」という表題の米国非仮特許出願(代理人整理番号ILLM 1000-5/IP-1611-US)。

【０００７】

同時に出願された、Laksshman Sundaram、Kai-How Farh、Hong Gao、およびJeremy Francis McRaeによる、「DEEP CONVOLUTIONAL NEURAL NETWORKS FOR VARIANT CLASSIFICATION」という表題の米国非仮特許出願(代理人整理番号ILLM 1000-6/IP-1612-US)。

【０００８】

同時に出願された、Laksshman Sundaram、Kai-How Farh、Hong Gao、およびJeremy Francis McRaeによる、「SEMI-SUPERVISED LEARNING FOR TRAINING AN ENSEMBLE OF DEEP CONVOLUTIONAL NEURAL NETWORKS」という表題の米国非仮特許出願(代理人整理番号ILLM 1000-7/IP-1613-US)。

【０００９】

文書1 - A.V.D.Oord、S.Dieleman、H.Zen、K.Simonyan、O.Vinyals、A.Graves、N.Kalchbrenner、A.Senior、およびK.Kavukcuoglu、「WAVENET: A GENERATIVE MODEL FOR RAW AUDIO」、arXiv:1609.03499、2016

【００１０】

文書2 - S.O.Arik、M.Chrzanowski、A.Coates、G.Diamos、A.Gibiansky、Y.Kang、X.Li、J.Miller、A.Ng、J.Raiman、S.Sengupta、およびM.Shoeybi、「DEEP VOICE: REAL-TIME NEURAL TEXT-TO-SPEECH」、arXiv:1702.07825、2017

【００１１】

文書3 - F.YuおよびV.Koltun、「MULTI-SCALE CONTEXT AGGREGATION BY DILATED CONVOLUTIONS」、arXiv:1511.07122、2016

【００１２】

文書4 - K.He、X.Zhang、S.Ren、およびJ.Sun、「DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION」、arXiv:1512.03385、2015

【００１３】

文書5 - R.K.Srivastava、K.Greff、およびJ.Schmidhuber、「HIGHWAY NETWORKS」、arXiv:1505.00387、2015

【００１４】

文書6 - G.Huang、Z.Liu、L.van der Maaten、およびK.Q.Weinberger、「DENSELY CONNECTED CONVOLUTIONAL NETWORKS」、arXiv:1608.06993、2017

【００１５】

文書7 - C.Szegedy、W.Liu、Y.Jia、P.Sermanet、S.Reed、D.Anguelov、D.Erhan、V.Vanhoucke、およびA.Rabinovich、「GOING DEEPER WITH CONVOLUTIONS」、arXiv:1409.4842、2014

【００１６】

文書8 - S.Ioffe、およびC.Szegedy、「BATCH NORMALIZATION: ACCELERATING DEEP NETWORK TRAINING BY REDUCING INTERNAL COVARIATE SHIFT」、arXiv:1502.03167、2015

【００１７】

文書9 - J.M.Wolterink、T.Leiner、M.A.Viergever、およびI.Isgum、「DILATED CONVOLUTIONAL NEURAL NETWORKS FOR CARDIOVASCULAR MR SEGMENTATION IN CONGENITAL HEART DISEASE」、arXiv:1704.03669、2017

【００１８】

文書10 - L.C.Piqueras、「AUTOREGRESSIVE MODEL BASED ON A DEEP CONVOLUTIONAL NEURAL NETWORK FOR AUDIO GENERATION」、Tampere University of Technology、2016

【００１９】

文書11 - J.Wu、「Introduction to Convolutional Neural Networks」、Nanjing University、2017

【００２０】

文書12 - I.J.Goodfellow、D.Warde-Farley、M.Mirza、A.Courville、およびY.Bengio、「CONVOLUTIONAL NETWORKS」、Deep Learning、MIT Press、2016

【００２１】

文書13 - J.Gu、Z.Wang、J.Kuen、L.Ma、A.Shahroudy、B.Shuai、T.Liu、X.Wang、およびG.Wang、「RECENT ADVANCES IN CONVOLUTIONAL NEURAL NETWORKS」、arXiv:1512.07108、2017

【００２２】

文書1は、入力シーケンスを受け入れて入力シーケンス中のエントリをスコアリングする出力シーケンスを生成するために、同じ畳み込みウィンドウサイズを有する畳み込みフィルタ、バッチ正規化層、正規化線形ユニット(ReLUと省略される)層、次元変換層、指数関数的に増大する膨張畳み込み率(atrous convolution rate)を伴う膨張畳み込み層、スキップ接続、およびソフトマックス分類層を伴う、残差ブロックのグループを使用する深層畳み込みニューラルネットワークアーキテクチャを説明する。開示される技術は、文書

1において説明されるニューラルネットワークコンポーネントおよびパラメータを使用する。一実装形態では、開示される技術は、文書1において説明されるニューラルネットワークコンポーネントのパラメータを修正する。たとえば、文書1とは異なり、開示される技術における膨張畳み込み率は、より低い残差ブロックグループからより高い残差ブロックグループへと非指数関数的に高まる。別の例では、文書1とは異なり、開示される技術における畳み込みウィンドウサイズは、残差ブロックのグループ間で変動する。

【００２３】

文書2は、文書1において説明される深層畳み込みニューラルネットワークアーキテクチャの詳細を説明する。

【００２４】

文書3は、開示される技術によって使用される膨張畳み込みを説明する。本明細書では、膨張畳み込みは「拡張畳み込み(dilated convolution)」とも呼ばれる。膨張/拡張畳み込みは、少数の訓練可能なパラメータで大きな受容野を可能にする。膨張/拡張畳み込みは、膨張畳み込み率または拡張係数とも呼ばれるあるステップを用いて入力値をスキップすることによって、カーネルがその長さより長いエリアにわたって適用されるような畳み込みである。膨張/拡張畳み込みは、畳み込み演算が実行されるときに、より長い間隔の隣り合う入力エントリ(たとえば、ヌクレオチド、アミノ酸)が考慮されるように、畳み込みフィルタ/カーネルの要素間に離隔を加える。これにより、入力における長距離のコンテクスト依存性の組み込みが可能になる。膨張畳み込みは、隣接するヌクレオチドが処理されるにつれて、部分的な畳み込み計算結果を再使用のために保存する。

【００２５】

文書4は、開示される技術によって使用される残差ブロックおよび残差接続を説明する。

【００２６】

文書5は、開示される技術によって使用されるスキップ接続を説明する。本明細書では、スキップ接続は「ハイウェイネットワーク」とも呼ばれる。

【００２７】

文書6は、開示される技術によって使用される密接続(densely connected)畳み込みネットワークアーキテクチャを説明する。

【００２８】

文書7は、開示される技術によって使用される次元変換畳み込み層およびモジュールベースの処理パイプラインを説明する。次元変換畳み込みの一例は1×1の畳み込みである。

【００２９】

文書8は、開示される技術によって使用されるバッチ正規化層を説明する。

【００３０】

文書9も、開示される技術によって使用される膨張/拡張畳み込みを説明する。

【００３１】

文書10は、畳み込みニューラルネットワーク、深層畳み込みニューラルネットワーク、および膨張/拡張畳み込みを伴う深層畳み込みニューラルネットワークを含む、開示される技術によって使用され得る深層ニューラルネットワークの様々なアーキテクチャを説明する。

【００３２】

文書11は、サブサンプリング層(たとえば、プーリング)および全結合層を伴う畳み込みニューラルネットワークを訓練するためのアルゴリズムを含む、開示される技術によって使用され得る畳み込みニューラルネットワークの詳細を説明する。

【００３３】

文書12は、開示される技術によって使用され得る様々な畳み込み演算の詳細を説明する。

【００３４】

文書13は、開示される技術によって使用され得る畳み込みニューラルネットワークの様

々なアーキテクチャを説明する。
【００３５】
本出願とともに電子的に提出されるテーブルの参照による引用
ASCIIテキストフォーマットの以下のテーブルファイルが、本明細書とともに提出され、参照によって引用される。ファイルの名称、作成日、およびサイズは次の通りである。
【００３６】
SupplementaryTable1.txt 2018年10月2日 13KB
【００３７】
SupplementaryTable2.txt 2018年10月2日 13KB
【００３８】
SupplementaryTable3.txt 2018年10月2日 11KB
【００３９】
SupplementaryTable4.txt 2018年10月2日 13KB
【００４０】
SupplementaryTable6.txt 2018年10月2日 12KB
【００４１】
SupplementaryTable7.txt 2018年10月2日 44KB
【００４２】
SupplementaryTable13.txt 2018年10月2日 119KB
【００４３】
SupplementaryTable18.txt 2018年10月2日 35KB
【００４４】
SupplementaryTable20.txt 2018年10月2日 1027KB
【００４５】
SupplementaryTable20Summary.txt 2018年10月2日 9KB
【００４６】
SupplementaryTable21.txt 2018年10月2日 24KB
【００４７】
SupplementaryTable21.txt 2018年10月2日 24KB
【００４８】
SupplementaryTable18.txt 2018年10月4日 35KB
【００４９】
DataFileS1.txt 2018年10月4日 138MB
【００５０】
DataFileS2.txt 2018年10月4日 980MB
【００５１】
DataFileS3.txt 2018年10月4日 1.01MB
【００５２】
DataFileS4.txt 2018年10月4日 834KB
【００５３】
Pathogenicity_prediction_model.txt 2018年10月4日 8.24KB
【００５４】
補足テーブル1:分析において使用される各種からのバリアントの詳細。このテーブルは、これらのデータソースの各々のためのパイプラインにおける中間結果を含む。このテーブルはSupplementaryTable1.txtにおいて提供されることに留意されたい。
【００５５】
補足テーブル2:一般的なヒトアレル頻度で他の種において存在するミスセンスバリアントの枯渇率。この枯渇率は、ヒトと他の種との間で同一状態(identical-by-state)であったバリアントを使用して、稀なバリアント(<0.1%)と比較された一般的なバリアント(>0.1%)におけるミスセンス:同義比に基づいて計算された。このテーブルはSupplementaryTabl

e2.txtにおいて提供されることに留意されたい。

【００５６】

補足テーブル3:ヒトと他の哺乳類との間で50%を超える平均ヌクレオチド保存率を伴う遺伝子だけに制約された、一般的なヒトアレル頻度で他の種において存在するミスセンスバリアントの枯渇率。この枯渇率は、ヒトと他の種との間で同一状態であったバリアントを使用して、稀なバリアント(<0.1%)と比較された一般的なバリアント(>0.1%)におけるミスセンス:同義比に基づいて計算された。このテーブルはSupplementaryTable3.txtにおいて提供されることに留意されたい。

【００５７】

補足テーブル4:一般的なヒトアレル頻度で関連する種のペアにおいて固定された置換として存在するミスセンスバリアントの枯渇率。この枯渇率は、ヒトと関連する種のペアとの間で同一状態であったバリアントを使用して、稀なバリアント(<0.1%)と比較された一般的なバリアント(>0.1%)におけるミスセンス:同義比に基づいて計算された。このテーブルはSupplementaryTable4.txtにおいて提供されることに留意されたい。

【００５８】

補足テーブル6:SCN2A遺伝子のドメイン固有のアノテーション。ウィルコクソンの順位和のp値は、タンパク質全体と比較した特定のドメインにおけるPrimateAIスコアの相違を示す。太字で強調されたドメインはタンパク質の約7%をカバーするが、ClinVar病原性アノテーションの大半を有する。このことは、それらのドメインに対する平均PrimateAIスコアとよく符合し、PrimateAIモデルによれば上位3つの病原性ドメインである。このテーブルはSupplementaryTable6.txtにおいて提供されることに留意されたい。

【００５９】

補足テーブル7:予想されるミスセンス:同義比に対するアレル頻度の影響を計算する際に使用される生カウント。同義バリアントとミスセンスバリアントの予想されるカウントは、変異率および遺伝子変換を考慮するためにトリヌクレオチドコンテクストを使用して、イントロン領域におけるバリアントに基づいて計算された。このテーブルはSupplementaryTables.xlsxにおいて提供されることに留意されたい。

【００６０】

補足テーブル13:3状態の二次構造および3状態の溶媒接触性予測のための深層学習モデルを訓練するために使用されるProtein DataBank(PDB)からのタンパク質名の一覧。ラベル列は、モデル訓練の訓練/妥当性確認/検定段階においてタンパク質が使用されるかどうかを示す。このテーブルはSupplementaryTable13.txtにおいて提供されることに留意されたい。

【００６１】

補足テーブル18:タンパク質切断変異(p<0.05)のみから計算される、DDD研究において疾患との関連について名目上有意であった605個の遺伝子の一覧。このテーブルはSupplementaryTable18.txtにおいて提供されることに留意されたい。

【００６２】

補足テーブル20:少なくとも1つの観察されるDNMを伴うすべての遺伝子に対する、遺伝子ごとのde novo変異(DNM)のエンリッチメントの検定結果。すべてのDNMを含むときの、およびPrimateAIスコアが0.803より小さいミスセンスDNMを除去した後の、P値が与えられる。FDRで訂正されたP値が同様に与えられる。DDDコホートだけからの、および完全なメタ分析コホートからの、観察されるタンパク質切断(PTV)DNMとミスセンスDNMのカウントも含まれる。第1にすべてのミスセンスDNMを含むときの、および第2にPrimateAIスコアが0.803より小さいすべてのミスセンスDNMを除去した後の、観察され予測されるミスセンスDNMの同様のカウントも含まれる。このテーブルはSupplementaryTable20.txtおよびSupplementaryTable20Summary.txtにおいて提供されることに留意されたい。

【００６３】

補足テーブル21:FDR<0.1である遺伝子におけるde novo変異のエンリッチメントを検定した結果。一度はすべてのミスセンスde novo変異についての、およびもう一度は損害を

引き起こすミスセンス変異だけについての、観察されるタンパク質切断(PTV)de novo変異のカウントと、他のタンパク質変換de novo変異のカウントとが含まれる。低スコアのミスセンス箇所を除外した後のP値と比較した、すべてのミスセンスサイトを含むときのP値が与えられる。このテーブルはSupplementaryTable21.txtにおいて提供されることに留意されたい。

【００６４】

DataFileS1:他の種において存在するすべてのバリアントの一覧。「ClinVar有意性」列は、利用可能な矛盾しないClinVarアノテーションを含む。このテーブルはDataFileS1.txtにおいて提供されることに留意されたい。

【００６５】

DataFileS2:関連する種のペアからのすべての固定された置換の一覧。このテーブルはDataFileS2.txtにおいて提供されることに留意されたい。

【００６６】

DataFileS3:霊長類とIBSである保留された(withheld)良性検定バリアントの一覧。良性検定バリアントは、1つ以上の霊長類の種とIBSである一般的ではないヒトバリアントである。このテーブルはDataFileS3.txtにおいて提供されることに留意されたい。

【００６７】

DataFileS4:保留された良性検定バリアントと一致する、霊長類とIBSであるラベリングされていないバリアントの一覧。ラベリングされていないバリアントは、変異率、カバレッジの偏り、および霊長類の種とのアラインメント可能性について、良性検定バリアントと一致する。このテーブルはDataFileS4.txtにおいて提供されることに留意されたい。

【００６８】

Pathogenicity_prediction_model:一実装形態に従って開示される技術を可能にするPythonプログラミング言語のコード。このコードファイルはPathogenicity_prediction_model.txtにおいて提供されることに留意されたい。

【００６９】

開示される技術の分野

開示される技術は、人工知能タイプコンピュータならびにデジタルデータ処理システムならびに知性のエミュレーションのための対応するデータ処理方法および製品(すなわち、知識ベースシステム、推論システム、知識取得システム)に関し、不確実性を伴う推論のためのシステム(たとえば、ファジー論理システム)、適応システム、機械学習システム、および人工ニューラルネットワークを含む。具体的には、開示される技術は、深層畳み込み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法を使用することに関する。

【背景技術】

【００７０】

このセクションにおいて論じられる主題は、このセクションにおける言及の結果として、単なる従来技術であると見なされるべきではない。同様に、このセクションにおいて言及される問題、または背景として提供される主題と関連付けられる問題は、従来技術においてこれまで認識されていたと見なされるべきではない。このセクションの主題は異なる手法を表すにすぎず、それらの異なる手法自体も、特許請求される技術の実装形態に対応し得る。

【００７１】

［機械学習］

機械学習では、出力変数を予測するために入力変数が使用される。入力変数はしばしば特徴量と呼ばれ、$X=(X_1,X_2,...,X_k)$と表記され、$i\in 1,...,k$である各X_iが特徴量である。出力変数はしばしば応答または依存変数と呼ばれ、変数Y_iにより表記される。Yと対応するXとの関係は、次の一般的な形式で書くことができる。

$Y=f(x)+\in$

【００７２】

上式において、fは特徴量($X_1,X_2,...,X_k$)の関数であり、∈はランダムな誤差の項である。この誤差の項は、Xとは無関係であり、平均値が0である。

【００７３】

実際には、特徴量Xは、Yがなくても、またはXとYとの厳密な関係を知らなくても入手可能である。誤差の項は平均値が0であるので、目標はfを推定することである。

【００７４】

【数１】

$$\hat{Y}=\hat{f}=(X)$$

【００７５】

上式において、

【００７６】

【数２】

$$\hat{f}$$

【００７７】

は∈の推定値であり、これはしばしばブラックボックスと見なされ、

【００７８】

【数３】

$$\hat{f}$$

【００７９】

の入力と出力の関係のみが知られていることを意味するが、なぜこれで機能するのかという疑問は答えられていないままである。

【００８０】

関数

【００８１】

【数４】

$$\hat{f}$$

【００８２】

は学習を使用して発見される。教師あり学習および教師なし学習は、このタスクのための機械学習において使用される2つの方式である。教師あり学習では、ラベリングされたデータが訓練のために使用される。入力および対応する出力(=ラベル)を示すことによって、関数

【００８３】

【数５】

$$\hat{f}$$

【００８４】

は、出力を近似するように最適化される。教師なし学習では、目標はラベリングされていないデータから隠された構造を見つけることである。このアルゴリズムは、入力データについての正確さの尺度を持たず、これにより教師あり学習と区別される。

【００８５】

［ニューラルネットワーク］

図1Aは、複数の層を伴う全結合ニューラルネットワークの一実装形態を示す。ニューラルネットワークは、互いとの間でメッセージを交換する相互接続された人工ニューロン(たとえば、a_1、a_2、a_3)のシステムである。示されるニューラルネットワークは3つの入力を有し、2つのニューロンが隠れ層にあり、2つのニューロンが出力層にある。隠れ層は活性化関数f(・)を有し、出力層は活性化関数g(・)を有する。これらの接続は、適切に訓練されたネットワークが認識すべき画像を与えられると正しく応答するように、訓練プロセスの間に調整された数値的な重み(たとえば、w_{11}、w_{21}、w_{12}、w_{31}、w_{22}、w_{32}、v_{11}、v_{22})を有する。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出力を処理する。出力層は、隠れ層から出力を取り込み、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。ネットワークは、特徴検出ニューロンの複数の層を含む。各層は、前の層からの入力の異なる組合せに対応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における基本的な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するように、構築される。

【００８６】

遺伝学における深層学習の応用の概観は、以下の出版物において見出され得る。

・ T.Ching他、Opportunities And Obstacles For Deep Learning In Biology And Medicine、www.biorxiv.org:142760、2017

・ Angermueller C、Parnamaa T、Parts L、Stegle O、Deep Learning For Computational Biology. Mol Syst Biol. 2016;12:878

・ Park Y、Kellis M、2015 Deep Learning For Regulatory Genomics. Nat. Biotechnol. 33、825-826、(doi:10.1038/nbt.3313)

・ Min S、Lee B、およびYoon S、Deep Learning In Bioinformatics. Brief. Bioinform. bbw068 (2016)

・ Leung MK、Delong A、Alipanahi B他、Machine Learning In Genomic Medicine: A Review of Computational Problems and Data Sets、2016

・ Libbrecht MW、Noble WS、Machine Learning Applications In Genetics and Genomics. Nature Reviews Genetics 2015;16(6):321-32

【先行技術文献】

【特許文献】

【００８７】

【特許文献１】PCT特許出願第PCT/US2018/(未定)号(代理人整理番号ILLM 1000-9/IP-1612-PCT)

【特許文献２】国際特許出願第PCT/US2018/(未定)号(代理人整理番号第ILLM 1000-10/IP-1613-PCT)

【特許文献３】米国特許出願第(未定)号(代理人整理番号第ILLM 1000-5/IP-1611-US)

【特許文献４】米国特許出願第(未定)号(代理人整理番号第ILLM 1000-6/IP-1612-US)

【特許文献５】米国特許出願第(未定)号(代理人整理番号第ILLM 1000-7/IP-1613-US)

【特許文献６】国際特許出願公開第WO07010252号

【特許文献７】国際特許出願第PCTGB2007/003798号

【特許文献８】米国特許出願公開第2009/0088327号

【特許文献９】米国特許出願公開第2016/0085910号

【特許文献１０】米国特許出願公開第2013/0296175号

【特許文献１１】国際特許出願公開第WO 04/018497号

【特許文献１２】米国特許第7057026号

【特許文献１３】国際特許出願公開第WO 91/06678号

【特許文献１４】国際特許出願公開第WO 07/123744号

【特許文献１５】米国特許第7329492号

【特許文献１６】米国特許第7211414号

【特許文献１７】米国特許第7315019号

【特許文献１８】米国特許第7405281号
【特許文献１９】米国特許出願公開第2008/0108082号
【特許文献２０】米国特許第5641658号
【特許文献２１】米国特許出願公開第2002/0055100号
【特許文献２２】米国特許第7115400号
【特許文献２３】米国特許出願公開第2004/0096853号
【特許文献２４】米国特許出願公開第2004/0002090号
【特許文献２５】米国特許出願公開第2007/0128624号
【特許文献２６】米国特許出願公開第2008/0009420号
【特許文献２７】米国特許出願公開第2007/0099208A1号
【特許文献２８】米国特許出願公開第2007/0166705A1号
【特許文献２９】米国特許出願公開第2008/0280773A1号
【特許文献３０】米国特許出願第13/018255号
【特許文献３１】国際特許出願公開第WO 2014/142831号
【非特許文献】
【００８８】
【非特許文献１】A.V.D.Oord、S.Dieleman、H.Zen、K.Simonyan、O.Vinyals、A.Graves、N.Kalchbrenner、A.Senior、およびK.Kavukcuoglu、「WAVENET: A GENERATIVE MODEL FOR RAW AUDIO」、arXiv:1609.03499、2016
【非特許文献２】S.O.Arik、M.Chrzanowski、A.Coates、G.Diamos、A.Gibiansky、Y.Kang、X.Li、J.Miller、A.Ng、J.Raiman、S.Sengupta、およびM.Shoeybi、「DEEP VOICE: REAL-TIME NEURAL TEXT-TO-SPEECH」、arXiv:1702.07825、2017
【非特許文献３】F.YuおよびV.Koltun、「MULTI-SCALE CONTEXT AGGREGATION BY DILATED CONVOLUTIONS」、arXiv:1511.07122、2016
【非特許文献４】K.He、X.Zhang、S.Ren、およびJ.Sun、「DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION」、arXiv:1512.03385、2015
【非特許文献５】R.K.Srivastava、K.Greff、およびJ.Schmidhuber、「HIGHWAY NETWORKS」、arXiv:1505.00387、2015
【非特許文献６】G.Huang、Z.Liu、L.van der Maaten、およびK.Q.Weinberger、「DENSELY CONNECTED CONVOLUTIONAL NETWORKS」、arXiv:1608.06993、2017
【非特許文献７】C.Szegedy、W.Liu、Y.Jia、P.Sermanet、S.Reed、D.Anguelov、D.Erhan、V.Vanhoucke、およびA.Rabinovich、「GOING DEEPER WITH CONVOLUTIONS」、arXiv:1409.4842、2014
【非特許文献８】S.Ioffe、およびC.Szegedy、「BATCH NORMALIZATION: ACCELERATING DEEP NETWORK TRAINING BY REDUCING INTERNAL COVARIATE SHIFT」、arXiv:1502.03167、2015
【非特許文献９】J.M.Wolterink、T.Leiner、M.A.Viergever、およびI.Isgum、「DILATED CONVOLUTIONAL NEURAL NETWORKS FOR CARDIOVASCULAR MR SEGMENTATION IN CONGENITAL HEART DISEASE」、arXiv:1704.03669、2017
【非特許文献１０】L.C.Piqueras、「AUTOREGRESSIVE MODEL BASED ON A DEEP CONVOLUTIONAL NEURAL NETWORK FOR AUDIO GENERATION」、Tampere University of Technology、2016
【非特許文献１１】J.Wu、「Introduction to Convolutional Neural Networks」、Nanjing University、2017
【非特許文献１２】I.J.Goodfellow、D.Warde-Farley、M.Mirza、A.Courville、およびY.Bengio、「CONVOLUTIONAL NETWORKS」、Deep Learning、MIT Press、2016
【非特許文献１３】J.Gu、Z.Wang、J.Kuen、L.Ma、A.Shahroudy、B.Shuai、T.Liu、X.Wang、およびG.Wang、「RECENT ADVANCES IN CONVOLUTIONAL NEURAL NETWORKS」、arXiv:1512.07108、2017
【非特許文献１４】T.Ching他、Opportunities And Obstacles For Deep Learning In Bi

ology And Medicine、www.biorxiv.org:142760、2017
【非特許文献１５】Angermueller C、Parnamaa T、Parts L、Stegle O、Deep Learning For Computational Biology. Mol Syst Biol. 2016;12:878
【非特許文献１６】Park Y、Kellis M、2015 Deep Learning For Regulatory Genomics. Nat. Biotechnol. 33、825-826、(doi:10.1038/nbt.3313)
【非特許文献１７】Min S、Lee B、およびYoon S、Deep Learning In Bioinformatics. Brief. Bioinform. bbw068 (2016)
【非特許文献１８】Leung MK、Delong A、Alipanahi B他、Machine Learning In Genomic Medicine: A Review of Computational Problems and Data Sets、2016
【非特許文献１９】Libbrecht MW、Noble WS、Machine Learning Applications In Genetics and Genomics. Nature Reviews Genetics 2015;16(6):321-32
【非特許文献２０】K.He、X.Zhang、S.Ren、およびJ.Sun、「DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION」、arXiv:1512.03385、2015
【非特許文献２１】Bentley他、Nature 456:53-59(2008)
【非特許文献２２】Lizardi他、Nat.Genet.19:225-232(1998)
【非特許文献２３】Dunn, TamsenおよびBerry, GwennおよびEmig-Agius, DorotheaおよびJiang, YuおよびIyer, AnitaおよびUdar, NitinおよびStromberg, Michael、2017、Pisces: An Accurate and Versatile Single Sample Somatic and Germline Variant Caller、595-595、10.1145/3107411.3108203
【非特許文献２４】T Saunders, ChristopherおよびWong, WendyおよびSwamy, SajaniおよびBecq, JenniferおよびJ Murray, LisaおよびCheetham, Keira、2012、Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs、Bioinformatics (Oxford, England)、28、1811-7、10.1093/bioinformatics/bts271
【非特許文献２５】Kim, S、Scheffler, K、Halpern, A.L.、Bekritsky, M.A、Noh, E、Kallberg M、Chen, X、Beyter, D、Krusche, P、およびSaunders, C.T、2017、Strelka2: Fast and accurate variant calling for clinical sequencing applications
【非特許文献２６】Stromberg, MichaelおよびRoy, RajatおよびLajugie, JulienおよびJiang, YuおよびLi, HaochenおよびMargulies, Elliott、2017、Nirvana: Clinical Grade Variant Annotator、596-596、10.1145/3107411.3108204において説明される、Illumina Inc.

【発明の概要】
【課題を解決するための手段】
【００８９】
図面において、同様の参照文字は一般に様々な図全体で同様の部分を指す。また、図面は必ずしも縮尺通りではなく、代わりに、開示される技術の原理を示す際に一般に強調が行われる。以下の説明では、開示される技術の様々な実装形態が、以下の図面を参照して説明される。
【図面の簡単な説明】
【００９０】
【図１Ａ】複数の層を伴う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である。
【図１Ｂ】畳み込みニューラルネットワークの動作の一実装形態の図である。
【図１Ｃ】開示される技術の一実装形態による畳み込みニューラルネットワークを訓練するブロック図である。
【図１Ｄ】開示される技術の一実装形態によるサブサンプリング層(平均/最大プーリング)の一実装形態の図である。
【図１Ｅ】開示される技術の一実装形態によるReLU非線形層の一実装形態を示す図である。
【図１Ｆ】畳み込み層の2層の畳み込みの一実装形態を示す図である。
【図１Ｇ】特徴マップの追加を介して以前の情報ダウンストリームを再注入する残差接続

を示す図である。
【図１Ｈ】残差ブロックおよびスキップ接続の一実装形態を示す図である。
【図１Ｉ】バッチ正規化フォワードパスを示す図である。
【図１Ｊ】検定時のバッチ正規化変換を示す図である。
【図１Ｋ】バッチ正規化バックワードパスを示す図である。
【図１Ｌ】畳み込み層または密結合層の後と前のバッチ正規化層の使用を示す図である。
【図１Ｍ】1D畳み込みの一実装形態を示す図である。
【図１Ｎ】グローバル平均プーリング(GAP)がどのように機能するかを示す図である。
【図１Ｏ】拡張畳み込みを示す図である。
【図１Ｐ】積層(stacked)拡張畳み込みの一実装形態を示す図である。
【図１Ｑ】開示される技術を動作させることができる例示的なコンピューティング環境を示す図である。
【図２】本明細書で「PrimateAI」と呼ばれる、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である。
【図３】病原性分類のための深層学習ネットワークアーキテクチャであるPrimateAIの概略図である。
【図４Ａ】病原性予測深層学習モデルPrimateAIの例示的なモデルアーキテクチャの詳細を示す補足テーブル16である。
【図４Ｂ】病原性予測深層学習モデルPrimateAIの例示的なモデルアーキテクチャの詳細を示す補足テーブル16である。
【図４Ｃ】病原性予測深層学習モデルPrimateAIの例示的なモデルアーキテクチャの詳細を示す補足テーブル16である。
【図５】タンパク質の二次構造および溶媒接触性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である。
【図６】タンパク質の二次構造および溶媒接触性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である。
【図７Ａ】3状態二次構造予測深層学習(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補足テーブル11である。
【図７Ｂ】3状態二次構造予測深層学習(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補足テーブル11である。
【図８Ａ】3状態溶媒接触性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補足テーブル12である。
【図８Ｂ】3状態溶媒接触性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補足テーブル12である。
【図９】良性バリアントおよび病原性バリアントから基準タンパク質配列および代替タンパク質配列を生成することの一実装形態を示す図である。
【図１０】基準タンパク質配列と代替タンパク質配列をアラインメントすることの一実装形態を示す図である。
【図１１】位置特異的重み行列(PWMと省略される)または位置特異的スコアリング行列(PSSMと省略される)と呼ばれる、位置特異的頻度行列(PFMと省略される)を生成することの一実装形態を示す図である。
【図１２】二次構造および溶媒接触性サブネットワークの処理を示す図である。
【図１３】二次構造および溶媒接触性サブネットワークの処理を示す図である。
【図１４】二次構造および溶媒接触性サブネットワークの処理を示す図である。
【図１５】二次構造および溶媒接触性サブネットワークの処理を示す図である。
【図１６】バリアント病原性分類器の動作を示す図である。本明細書では、バリアントという用語は、一塩基多型(SNPと省略される)も指し、一般に一塩基バリアント(SNVと省略される)を指す。
【図１７】残差ブロックを示す図である。
【図１８】二次構造および溶媒接触性サブネットワークのニューラルネットワークアーキ

テクチャを示す図である。
【図１９】バリアント病原性分類器のニューラルネットワークアーキテクチャを示す図である。
【図２０】重要な機能ドメインのためにアノテートされた、SCN2A遺伝子の中の各アミノ酸の場所における予測される病原性スコアを示す図である。
【図２１Ｄ】訓練を保留された10000個の一般的な霊長類のバリアントの検定セットに対する良性の結果を予測することにおける分類器の比較を示す図である。
【図２１Ｅ】Deciphering Developmental Disorders(DDD)の患者において発生するde novoミスセンスバリアントに対するPrimateAI予測スコアの分布を、影響を受けていない兄弟と比較して、対応するウィルコクソンの順位和のP値とともに示す図である。
【図２１Ｆ】DDD症例群vs対照群におけるde novoミスセンスバリアントを分離する際の分類器の比較を示す図である。ウィルコクソンの順位和検定のP値が各分類器に対して示されている。
【図２２Ａ】de novoタンパク質切断変異(P<0.05)に対して有意であった605個の関連する遺伝子内での、DDDコホートからの影響を受けている個人における予想を超えるde novoミスセンス変異のエンリッチメントを示す図である。
【図２２Ｂ】605個の関連する遺伝子内での、DDD患者vs影響を受けていない兄弟において発生するde novoミスセンスバリアントに対するPrimateAI予測スコアの分布を、対応するウィルコクソンの順位和のP値とともに示す図である。
【図２２Ｃ】605個の遺伝子内での症例群vs対照群におけるde novoミスセンスバリアントを分離する際の様々な分類器の比較を示す図である。
【図２２Ｄ】各分類器に対して示される曲線下面積(AUC)とともに、受信者動作特性曲線上で示される、様々な分類器の比較を示す図である。
【図２２Ｅ】各分類器に対する分類の正確さおよび曲線下面積(AUC)を示す図である。
【図２３Ａ】訓練のために使用されるデータの分類の正確さに対する影響を示す図である。
【図２３Ｂ】訓練のために使用されるデータの分類の正確さに対する影響を示す図である。
【図２３Ｃ】訓練のために使用されるデータの分類の正確さに対する影響を示す図である。
【図２３Ｄ】訓練のために使用されるデータの分類の正確さに対する影響を示す図である。
【図２４】一般的な霊長類バリアントの確認に対するシーケンシングカバレッジの影響を訂正することを示す図である。
【図２５Ａ】開示されるニューラルネットワークによるタンパク質モチーフの認識を示す図である。
【図２５Ｂ】開示されるニューラルネットワークによるタンパク質モチーフの認識を示す図である。
【図２５Ｃ】開示されるニューラルネットワークによるタンパク質モチーフの認識を示す図である。
【図２６】開示されるニューラルネットワークによるタンパク質モチーフの認識を示す図である。バリアントに対する予測される深層学習スコアへの、バリアントの中および周りの各場所を摂動させることの影響を示す線プロットを含む。
【図２７】重みの相関パターンがBLOSUM62スコア行列およびGranthamスコア行列に倣っていることを示す図である。
【図２８Ａ】深層学習ネットワークのPrimateAIおよび他の分類器の性能評価を示す図である。
【図２８Ｂ】深層学習ネットワークのPrimateAIおよび他の分類器の性能評価を示す図である。
【図２８Ｃ】深層学習ネットワークのPrimateAIおよび他の分類器の性能評価を示す図で

ある。
【図２９Ａ】4つの分類器の予測スコアの分布を示す図である。
【図２９Ｂ】4つの分類器の予測スコアの分布を示す図である。
【図３０Ａ】605個の疾患関連遺伝子において病原性バリアントと良性バリアントとを分離する際のPrimateAIネットワークおよび他の分類器の正確さを比較する図である。
【図３０Ｂ】605個の疾患関連遺伝子において病原性バリアントと良性バリアントとを分離する際のPrimateAIネットワークおよび他の分類器の正確さを比較する図である。
【図３０Ｃ】605個の疾患関連遺伝子において病原性バリアントと良性バリアントとを分離する際のPrimateAIネットワークおよび他の分類器の正確さを比較する図である。
【図３１Ａ】専門家により精選されたClinVarバリアントに対する分類器の性能と、経験的なデータセットに対する性能との相関を示す図である。
【図３１Ｂ】専門家により精選されたClinVarバリアントに対する分類器の性能と、経験的なデータセットに対する性能との相関を示す図である。
【図３２】Protein Databankからのアノテートされたサンプルに対する3状態二次構造予測モデルおよび3状態溶媒接触性予測モデルの性能を示す補足テーブル14である。
【図３３】DSSPデータベースからのヒトタンパク質のアノテートされた二次構造ラベルを使用した深層学習ネットワークの性能比較を示す補足テーブル15である。
【図３４】評価した20個の分類器の各々に対する、10000個の保留された霊長類バリアントに対する正確さの値と、DDD症例群vs対照群におけるde novoバリアントに対するp値とを示す、補足テーブル17である。
【図３５】605個の疾患関連遺伝子に制約された、DDD症例群データセットvs対照群データセットにおけるde novoバリアントに対する異なる分類器の性能の比較を示す補足テーブル19である。
【図３６】開示される半教師あり学習器のコンピューティング環境を示す図である。
【図３７】開示される半教師あり学習の様々なサイクルを示す図である。
【図３８】開示される半教師あり学習の様々なサイクルを示す図である。
【図３９】開示される半教師あり学習の様々なサイクルを示す図である。
【図４０】開示される半教師あり学習の様々なサイクルを示す図である。
【図４１】開示される半教師あり学習の様々なサイクルを示す図である。
【図４２】反復的な均衡のとれたサンプリングプロセスを示す図である。
【図４３】良性データセットを生成するために使用されるコンピューティング環境の一実装形態を示す図である。
【図４４】良性ヒトミスセンスSNPを生成することの一実装形態を示す図である。
【図４５】ヒトオーソロガスミスセンスSNPの一実装形態を示す図である。ヒトと一致する基準コドンおよび代替コドンを有する、ヒト以外の種におけるミスセンスSNP。
【図４６】ヒトと一致する基準コドンを伴うヒト以外の霊長類の種(たとえば、チンパンジー)のSNPを良性として分類することの一実装形態を示す図である。
【図４７】エンリッチメントスコアを計算してそれらを比較することの一実装形態を示す図である。
【図４８】良性SNPデータセットの一実装形態を示す図である。
【図４９Ａ】ヒトアレル頻度スペクトラムにわたるミスセンス:同義比を示す図である。
【図４９Ｂ】ヒトアレル頻度スペクトラムにわたるミスセンス:同義比を示す図である。
【図４９Ｃ】ヒトアレル頻度スペクトラムにわたるミスセンス:同義比を示す図である。
【図４９Ｄ】ヒトアレル頻度スペクトラムにわたるミスセンス:同義比を示す図である。
【図４９Ｅ】ヒトアレル頻度スペクトラムにわたるミスセンス:同義比を示す図である。
【図５０Ａ】他の種と同一状態であるミスセンスバリアントに対する純化選択を示す図である。
【図５０Ｂ】他の種と同一状態であるミスセンスバリアントに対する純化選択を示す図である。
【図５０Ｃ】他の種と同一状態であるミスセンスバリアントに対する純化選択を示す図で

ある。
【図５０Ｄ】他の種と同一状態であるミスセンスバリアントに対する純化選択を示す図である。
【図５１】純化選択がない場合のヒトアレル頻度スペクトラムにわたる予想されるミスセンス:同義比を示す図である。
【図５２Ａ】CpGバリアントおよび非CpGバリアントに対するミスセンス:同義比を示す図である。
【図５２Ｂ】CpGバリアントおよび非CpGバリアントに対するミスセンス:同義比を示す図である。
【図５２Ｃ】CpGバリアントおよび非CpGバリアントに対するミスセンス:同義比を示す図である。
【図５２Ｄ】CpGバリアントおよび非CpGバリアントに対するミスセンス:同義比を示す図である。
【図５３】6種の霊長類と同一状態であるヒトバリアントのミスセンス:同義比を示す図である。
【図５４】6種の霊長類と同一状態であるヒトバリアントのミスセンス:同義比を示す図である。
【図５５】6種の霊長類と同一状態であるヒトバリアントのミスセンス:同義比を示す図である。
【図５６】調査されたヒトコホートのサイズを増やすことによって発見された新しい一般的なミスセンスバリアントの飽和を示すシミュレーションである。
【図５７】ゲノムにおける異なる保存プロファイルにわたるPrimateAIの正確さを示す図である。
【図５８】一般的なヒトバリアントおよびヒト以外の霊長類において存在するバリアントからのラベリングされた良性訓練データセットへの寄与を示す補足テーブル5である。
【図５９】予想されるミスセンス:同義比に対するアレル頻度の影響を示す補足テーブル8である。
【図６０】ClinVar分析を示す補足テーブル9である。
【図６１】一実装形態による、ClinVarにおいて見出される他の種からのミスセンスバリアントの数を示す補足テーブル10である。
【図６２】知的障害における14個の追加の遺伝子候補の発見の一実装形態を示すテーブル1である。
【図６３】ClinVarにおける病原性バリアントと良性バリアントとの間のGranthamスコアの平均の差の一実装形態を示すテーブル2である。
【図６４】遺伝子ごとのエンリッチメント分析の一実装形態を示す図である。
【図６５】ゲノムワイドエンリッチメント分析の一実装形態を示す図である。
【図６６】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化されたブロック図である。
【発明を実施するための形態】
【００９１】
以下の議論は、あらゆる当業者が開示される技術を作成して使用することを可能にするために提示され、特定の適用例およびその要件の文脈で与えられる。開示される実装形態への様々な修正が当業者に容易に明らかとなり、本明細書で定義される一般的な原理は、開示される技術の趣旨および範囲から逸脱することなく他の実装形態および適用例に適用され得る。したがって、開示される技術は、示される実装形態に限定されることは意図されず、本明細書で開示される原理および特徴と矛盾しない最も広い範囲を認められるべきである。
【００９２】
［導入］
［畳み込みニューラルネットワーク］

畳み込みニューラルネットワークは特別なタイプのニューラルネットワークである。密結合層と畳み込み層との間の基本的な違いは、密層が入力特徴空間におけるグローバルパターンを学習するのに対して、畳み込み層がローカルパターンを学習するということである。画像の場合、入力の小さい2Dウィンドウにおいてパターンが見出される。この重要な特徴は、(1)畳み込みニューラルネットワークの学習するパターンが移動不変である、および(2)畳み込みニューラルネットワークがパターンの空間的階層を学習できるという、2つの興味深い特性を畳み込みニューラルネットワークに与える。

【００９３】

第1の特性に関して、写真の右下の角のあるパターンを学習した後、畳み込み層はそれをどこでも、たとえば左上の角において認識することができる。密結合ネットワークは、パターンが新しい位置において現れた場合、改めてパターンを学習しなければならない。これにより、畳み込みニューラルネットワークはデータ効率が高くなり、それは、一般化能力を有する表現を学習するのにより少数の訓練サンプルしか必要としないからである。

【００９４】

第2の特性に関して、第1の畳み込み層は端などの小さいローカルパターンを学習することができ、第2の畳み込み層は第1の層の特徴から作られるより大きいパターンを学習し、以下同様である。これにより、畳み込みニューラルネットワークは、ますます複雑になり抽象的になる視覚的な概念を効率的に学習することが可能になる。

【００９５】

畳み込みニューラルネットワークは、多くの異なる層において配置される人工ニューロンの層を、それらの層を互いに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、高度に非線形なマッピングを学習する。畳み込みニューラルネットワークは、1つまたは複数のサブサンプリング層および非線形層とともに散在する、1つまたは複数の畳み込み層を含み、サブサンプリング層および非線形層の後には、通常は1つまたは複数の全結合層がある。畳み込みニューラルネットワークの各要素は、以前の層における特徴のセットから入力を受け取る。畳み込みニューラルネットワークは同時に学習し、それは同じ特徴マップの中のニューロンが同一の重みを有するからである。これらの局所の共有される重みがネットワークの複雑さを下げるので、多次元入力データがネットワークに入るとき、畳み込みニューラルネットワークは、特徴の抽出および回帰または分類のプロセスにおいて、データ再構築の複雑さを避ける。

【００９６】

畳み込みは、2つの空間軸(高さおよび幅)ならびに深さ軸(チャネル軸とも呼ばれる)を伴う、特徴マップと呼ばれる3Dテンソルにわたって行われる。RGB画像では、深さ軸の次元は3であり、それは画像が3つの色チャネル、すなわち赤、緑、および青を有するからである。白黒の写真では、深さは1(グレーのレベル)である。畳み込み演算は、入力特徴マップからパッチを抽出し、これらのパッチのすべてに同じ変換を適用し、出力特徴マップを生成する。この出力特徴マップはそれでも3Dテンソルであり、幅および高さを有する。その深さは任意であってよく、それは出力深さが層のパラメータであり、その深さ軸における異なるチャネルはRGB入力におけるような特定の色をもはや表さず、むしろフィルタを表すからである。フィルタは入力データの特定の態様を符号化し、高いレベルで、単一のフィルタが、たとえば「入力における顔の存在」という概念を符号化することができる。

【００９７】

たとえば、第1の畳み込み層は、サイズ(28,28,1)の特徴マップを取り込み、サイズ(26,26,32)の特徴マップを出力する。すなわち、第1の畳み込み層は、その入力にわたる32個のフィルタを計算する。これらの32個の出力チャネルの各々が26×26の値の格子を含み、この格子は入力にわたるフィルタの応答マップであり、入力の中の異なる位置におけるそのフィルタパターンの応答を示す。これが、特徴マップという用語が意味することである。すなわち、深さ軸におけるそれぞれの次元が特徴(またはフィルタ)であり、2Dテンソル出力[:,:,n]が入力にわたるこのフィルタの応答の2D空間マップである。

【００９８】
畳み込みは、(1)通常は1×1、3×3、または5×5である入力から抽出されたパッチのサイズ、および(2)出力特徴マップの深さという、2つの重要なパラメータによって定義され、フィルタの数は畳み込みによって計算される。しばしば、これらは32という深さで開始し、64という深さまで続き、128または256という深さで終わる。
【００９９】
畳み込みは、3D入力特徴マップにわたってサイズ3×3または5×5のこれらのウィンドウをスライドし、それぞれの位置において止まり、周囲の特徴の3Dパッチ(形状(window_height、window_width、input_depth))を抽出することによって機能する。各々のそのような3Dパッチは次いで、形状の1Dベクトル(output_depth)への(畳み込みカーネルと呼ばれる、同じ学習された重み行列を伴うテンソル積を介して)変換される。これらのベクトルのすべてが次いで、形状の3D出力マップ(高さ、幅、output_depth)へと空間的に再び組み立てられる。出力特徴マップの中のそれぞれの空間的位置が入力特徴マップの中の同じ位置に対応する(たとえば、出力の右下の角は入力の右下の角についての情報を含む)。たとえば、3×3のウィンドウでは、ベクトル出力[i,j,:]は3Dパッチ入力[i-1:i+1,j-1:J+1,:]から来る。完全なプロセスは図1Bにおいて詳述される。
【０１００】
畳み込みニューラルネットワークは、訓練の間に多数の勾配更新反復を介して学習される入力値と畳み込みフィルタ(重みの行列)との間で畳み込み演算を実行する、畳み込み層を備える。(m,n)をフィルタサイズとし、Wは重みの行列とすると、畳み込み層は、ドット積w・x+bを計算することによって、入力Xを用いてWの畳み込みを実行し、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。同じ畳み込みフィルタが入力の異なる場所にわたって適用され、このことは学習される重みの数を減らす。このことは、すなわち、重要なパターンが入力において存在する場合、位置不変学習も可能にし、畳み込みフィルタは、重要なパターンがシーケンスの中でどこにあるかにかかわらず、重要なパターンを学習する。
【０１０１】
［畳み込みニューラルネットワークの訓練］
図1Cは、開示される技術の一実装形態による畳み込みニューラルネットワークを訓練することのブロック図を示す。畳み込みニューラルネットワークは、入力データが特定の出力推定につながるように、調整または訓練される。畳み込みニューラルネットワークは、出力推定とグラウンドトゥルースの比較に基づいて、出力推定がグラウンドトゥルースに漸近的に一致または接近するまで、逆伝播を使用して調整される。
【０１０２】
畳み込みニューラルネットワークは、グラウンドトゥルースと実際の出力との間の差に基づいてニューロン間の重みを調整することよって訓練される。これは次のように数学的に表される。
【０１０３】
【数６】

$$\Delta w_i = x_i \delta$$

【０１０４】
ただし、δ=(グラウンドトゥルース)-(実際の出力)
【０１０５】
一実装形態では、訓練規則は次のように定義される。
$w_{nm} \leftarrow w_{nm} + \alpha (t_m - \phi_m) \alpha_n$
【０１０６】
上式において、矢印は値の更新を示し、t_mはニューロンmの目標値であり、ϕ_mはニュー

ロンmの計算された現在の出力であり、α_nは入力nであり、αは学習率である。

【０１０７】

訓練における中間ステップは、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。出力において開始して、各層における重みに関する勾配が計算される。これは、バックワードパス、または後ろに行くと呼ばれる。ネットワークにおける重みは、負の勾配および以前の重みの組合せを使用して更新される。

【０１０８】

一実装形態では、畳み込みニューラルネットワークは、勾配降下法によって誤差の逆伝播を実行する確率的勾配更新アルゴリズム(ADAMなど)を使用する。シグモイド関数ベースの逆伝播アルゴリズムの一例は以下のように記述される。

【０１０９】

【数７】

$$\varphi = f(h) = \frac{1}{1+e^{-h}}$$

【０１１０】

上のシグモイド関数において、hはニューロンによって計算される加重和である。シグモイド関数は以下の導関数を有する。

【０１１１】

【数８】

$$\frac{\partial \varphi}{\partial h} = \varphi(1-\varphi)$$

【０１１２】

このアルゴリズムは、ネットワークの中のすべてのニューロンの活性化を計算し、フォワードパスに対する出力を生み出すことを含む。隠れ層の中のニューロンmの活性化は次のように記述される。

【０１１３】

【数９】

$$\varphi_m = \frac{1}{1+e^{-hm}}$$

$$h_m = \sum_{n=1}^{N} a_n w_{nm}$$

【０１１４】

これは、次のように記述される活性化を得るためにすべての隠れ層に対して行われる。

【０１１５】

【数１０】

$$\varphi_k = \frac{1}{1+e^{hk}}$$

$$h_k = \sum_{m=1}^{M} \varphi_m v_{mk}$$

【０１１６】
そして、誤差および訂正重みが層ごとに計算される。出力における誤差は次のように計算される。
$\delta_{ok}=(t_k-\varphi_k)\varphi_k(1-\varphi_k)$
【０１１７】
隠れ層における誤差は次のように計算される。
【０１１８】
【数１１】

$$\delta_{hm}=\varphi_m(1-\varphi_m)\sum_{k=1}^{K}v_{mk}\delta_{ok}$$

【０１１９】
出力層の重みは次のように更新される。
$v_{mk}\leftarrow v_{mk}+\alpha\ \delta_{ok}\varphi_m$
【０１２０】
隠れ層の重みは学習率αを使用して次のように更新される。
$v_{nm}\leftarrow w_{nm}+\alpha\ \delta_{hm}a_n$
【０１２１】
一実装形態では、畳み込みニューラルネットワークは、すべての層にわたって誤差を計算するために勾配降下最適化を使用する。そのような最適化において、入力特徴ベクトルxおよび予測される出力
【０１２２】
【数１２】

$$\hat{y}$$

【０１２３】
に対して、目標がyであるときに
【０１２４】
【数１３】

$$\hat{y}$$

【０１２５】
を予測することのコストのためのlとして損失関数が定義され、すなわち
【０１２６】
【数１４】

$$l(\hat{y},y)$$

【０１２７】
である。予測される出力
【０１２８】
【数１５】

$$\hat{y}$$

【０１２９】
は、関数fを使用して入力特徴ベクトルxから変換される。関数fは、畳み込みニューラル

ネットワークの重みによってパラメータ化され、すなわち

【０１３０】

【数１６】

$$\hat{y} = f_w(x)$$

【０１３１】

である。損失関数は

【０１３２】

【数１７】

$$l(\hat{y}, y) = l(f_w(x), y)$$

【０１３３】

、またはQ(z,w)=l(f$_w$(x),y)と記述され、ここでzは入力データと出力データのペア(x,y)である。勾配降下最適化は、以下に従って重みを更新することによって実行される。

【０１３４】

【数１８】

$$v_{t+1} = \mu v_t - \alpha \frac{1}{n} \sum_{i=1}^{N} \nabla w_t Q(z_t, w_t)$$

【０１３５】

$w_{t+1}=w_t+v_{t+1}$

【０１３６】

上式において、αは学習率である。また、損失はn個のデータペアのセットにわたる平均として計算される。この計算は、線形収束の際に学習率αが十分小さくなると終了する。他の実装形態では、計算効率をもたらすために、ネステロフの加速勾配法および適応勾配法に供給される選択されたデータペアだけを使用して、勾配が計算される。

【０１３７】

一実装形態では、畳み込みニューラルネットワークは、コスト関数を計算するために確率的勾配降下法(SGD)を使用する。SGDは、損失関数における重みに関する勾配を、以下で記述されるように、1つのランダム化されたデータペアz_tだけから計算することによって近似する。

$v_{t+1}=\mu v-\alpha \nabla wQ(z_t,w_t)$

$w_{t+1}=w_t+v_{t+1}$

【０１３８】

上式において、αは学習率であり、μはモメンタムであり、tは更新前の現在の重み状態である。SGDの収束速度は、学習率αが十分に速く低減するときと、十分に遅く低減するときの両方において、約O(1/t)である。他の実装形態では、畳み込みニューラルネットワークは、ユークリッド損失およ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。さらなる実装形態では、Adam確率的最適化器が畳み込みニューラルネットワークによって使用される。

【０１３９】

［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。畳み込み層は、入力データを学習して階層的特徴へと分解することが可能な、適応特徴抽出器として活動する。一実装形態では、畳み込み層は、入力として2つの画像を取り込み、出力として第3の画像を生成する。そのような実装形態では、畳み込みは2次元(2D)において2つの画像に対して動作し、一方の画像が入力画像であり、「カーネル」と呼ばれる他方の

画像が入力画像に対してフィルタとして適用され、出力画像を生成する。したがって、長さnの入力ベクトルfおよび長さmのカーネルgに対して、fとgの畳み込みf*gは次のように定義される。

【０１４０】

【数１９】

$$(f*g)(i)=\sum_{j=1}^{m} g(j)\cdot f(i-j+m/2)$$

【０１４１】

畳み込み演算は、入力画像にわたってカーネルをスライドすることを含む。カーネルの各場所に対して、カーネルと入力画像の重複する値が乗算され、結果が加算される。この積の合計が、カーネルが中心とされる入力画像の中の点における出力画像の値である。多数のカーネルから得られた異なる出力が特徴マップと呼ばれる。

【０１４２】

畳み込み層が訓練されると、それらは新しい推論データに対する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するので、明示的な特徴抽出を避け、訓練データから暗黙的に学習する。畳み込み層は畳み込みフィルタカーネル重みを使用し、これは訓練プロセスの一部として決定され更新される。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において組み合わされる。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し、それらの各々が、カーネルサイズ、ストライド、パディング、特徴マップの数、および重みなどの異なる畳み込みパラメータを伴う。

【０１４３】

［サブサンプリング層］

図1Dは、開示される技術の一実装形態によるサブサンプリング層の一実装形態である。サブサンプリング層は、抽出された特徴または特徴マップをノイズおよび歪みに対してロバストにするために、畳み込み層によって抽出される特徴の分解能を下げる。一実装形態では、サブサンプリング層は、2つのタイプのプーリング動作、すなわち平均プーリングおよび最大プーリングを利用する。プーリング動作は、入力を重複しない2次元空間へと分割する。平均プーリングでは、領域の中の4つの値の平均が計算される。最大プーリングでは、4つの値の最大値が選択される。

【０１４４】

一実装形態では、サブサンプリング層は、その出力を最大プーリングにおける入力のうちの1つだけにマッピングし、そ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る、以前の層の中のニューロンのセットに対するプーリング動作を含む。最大プーリングにおいて、プーリングニューロンの出力は、

$\Phi_0=\max(\Phi_1,\Phi_2,\ldots,\Phi_N)$

により記述されるような、入力の中に存在する最大値である。

【０１４５】

上式において、Nはニューロンセット内の要素の総数である。

【０１４６】

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、

【０１４７】

【数２０】

$$\varphi_o=\frac{1}{N}\sum_{n=1}^{N}\varphi_n$$

【０１４８】

によって記述されるような、入力ニューロンセットとともに存在する入力値の平均値であ

る。

【０１４９】

上式において、Nは入力ニューロンセット内の要素の総数である。

【０１５０】

図1Dにおいて、入力は4×4のサイズである。2×2のサブサンプリングに対して、4×4の画像は2×2のサイズの4つの重複しない行列へと分割される。平均プーリングでは、4つの値の平均は全整数出力である。最大プーリングでは、2×2の行列の中の4つの値の最大値は全整数出力である。

【０１５１】

［非線形層］

図1Eは、開示される技術の一実装形態による、非線形層の一実装形態を示す。非線形層は、各隠れ層上の可能性の高い特徴の明確な識別情報をシグナリングするために、異なる非線形トリガ関数を使用する。非線形層は、正規化線形ユニット(ReLU)、双曲線正接、双曲線正接の絶対値、シグモイドおよび連続トリガ(非線形)関数を含む、非線形トリガリングを実施するために様々な固有の関数を使用する。一実装形態では、ReLU活性化は、関数y=max(x,0)を実装し、層の入力サイズおよび出力サイズを同じに保つ。ReLUを使用することの利点は、畳み込みニューラルネットワークがより高速に多くの回数訓練されることである。ReLUは、入力が0以上の場合には、入力に関して線形であり、それ以外の場合には0である、非連続で非飽和の活性化関数である。数学的には、ReLU活性化関数は次のように記述される。

$\varphi(h)=\max(h,0)$

【０１５２】

【数２１】

$$\varphi(h)=\begin{cases} h & if\ h>0 \\ 0 & if\ h\leq 0 \end{cases}$$

【０１５３】

他の実装形態では、畳み込みニューラルネットワークは、

$\varphi(h)=(a+bh)^c$

によって記述される連続的な非飽和の関数である、冪ユニット活性化関数を使用する。

【０１５４】

上式において、a、b、およびcはそれぞれ、シフト、スケール、および冪を制御するパラメータである。冪活性化関数は、cが奇数の場合にはxとyで非対称な活性化を生み出し、cが偶数の場合にはy対称な活性化を生み出すことが可能である。いくつかの実装形態では、このユニットは非正規化線形活性化を生み出す。

【０１５５】

さら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、以下のロジスティック関数

【０１５６】

【数２２】

$$\varphi(h)=\frac{1}{1+e^{-\beta h}}$$

【０１５７】

によって記述される、連続的な飽和する関数である、シグモイドユニット活性化関数を使用する。

【０１５８】

上式において、β=1である。シグモイドユニット活性化関数は、負の活性化を生み出さ

ず、y軸に関してのみ非対称である。

【０１５９】

［畳み込みの例］

図1Fは、畳み込み層の2層の畳み込みの一実装形態を示す。図1Fにおいて、2048次元のサイズの入力が畳み込まれる。畳み込み1において、入力はサイズ3×3の16個のカーネルの2つのチャネルからなる畳み込み層によって畳み込まれる。得られる16個の特徴マップが次いで、ReLU1におけるReLU活性化関数によって正規化され、次いでサイズ3×3のカーネルを伴う16個のチャネルプーリング層を使用して平均プーリングによってプール1においてプールされる。畳み込み2において、プール1の出力が次いで、3×3のサイズを伴う30個のカーネルの16個のチャネルからなる別の畳み込み層によって畳み込まれる。さらに別のReLU2および2×2のカーネルサイズを伴うプール2における平均プーリングが、それに続く。畳み込み層は、可変の数、たとえば0個、1個、2個、および3個の、ストライドおよびパディングを使用する。得られる特徴ベクトルは、一実装形態によれば、512次元である。

【０１６０】

他の実装形態では、畳み込みニューラルネットワークは、異なる数の畳み込み層、サブサンプリング層、非線形層、および全結合層を使用する。一実装形態では、畳み込みニューラルネットワークは、より少数の層および層当たりのより多数のニューロンを伴う浅いネットワークであり、たとえば、層当たり100個から200個のニューロンを伴う、1個、2個、または3個の全結合層である。別の実装形態では、畳み込みニューラルネットワークは、より多数の層および層当たりのより少数のニューロンを伴う深層ネットワークであり、たとえば、層当たり30個から50個のニューロンを伴う、5個、6個、または8個の全結合層である。

【０１６１】

［フォワードパス］

特徴マップの中のf個の畳み込みコアに対するl番目の畳み込み層およびk番目の特徴マップにおける行x、列yのニューロンの出力は、次の式によって決定される。

【０１６２】

【数２３】

$$O_{x,y}^{(l,k)} = \tanh\left(\sum_{t=0}^{f-1}\sum_{r=0}^{k_h}\sum_{c=0}^{k_w} W_{(r,c)}^{(k,t)} O_{(x+r,x+c)}^{(l-1,t)} + Bias^{(l,k)}\right)$$

【０１６３】

l番目のサブサンプル層およびk番目の特徴マップにおける行x、列yのニューロンの出力は、次の式によって決定される。

【０１６４】

【数２４】

$$O_{x,y}^{(l,k)} = \tanh\left(W^{(k)}\sum_{r=0}^{S_h}\sum_{c=0}^{S_w} O_{(x\times S_h+r,y\times S_w+c)}^{(l-1,k)} + Bias^{(l,k)}\right)$$

【０１６５】

l番目の出力層のi番目のニューロンの出力は、次の式によって決定される。

【０１６６】

【数２５】

$$O_{(l,i)} = \tanh\left(\sum_{j=0}^{H} O_{(l-1,j)} W_{(i,j)}^{l} + Bias^{(l,i)}\right)$$

【０１６７】
［逆伝播］
出力層の中のk番目のニューロンの出力偏差は、次の式によって決定される。
【０１６８】
【数２６】

$$d(O_k^o) = y_k - t_k$$

【０１６９】
出力層の中のk番目のニューロンの入力偏差は、次の式によって決定される。
【０１７０】
【数２７】

$$d(I_k^o) = (y_k - t_k)\varphi'(v_k) = \varphi'(v_k)d(O_k^o)$$

【０１７１】
出力層の中のk番目のニューロンの重みおよびバイアスのばらつきは、次の式によって決定される。
【０１７２】
【数２８】

$$\Delta W_{k,x}^o) = d(I_k^o)y_{k,x}$$

$$\Delta Bias_k^o) = d(I_k^o)$$

【０１７３】
隠れ層の中のk番目のニューロンの出力バイアスは、次の式によって決定される。
【０１７４】
【数２９】

$$d(O_k^H) = \sum_{i=0}^{i<84} d(I_i^o)W_{i,k}$$

【０１７５】
隠れ層の中のk番目のニューロンの入力バイアスは、次の式によって決定される。
【０１７６】
【数３０】

$$d(I_k^H) = \varphi'(v_k)d(O_k^H)$$

【０１７７】
隠れ層の中のk個のニューロンから入力を受け取る前の層のm番目の特徴マップの中の行x、列yにおける重みおよびバイアスのばらつきは、次の式によって決定される。
【０１７８】

【数３１】

$$\Delta W_{m,x,y}^{H,k}) = d(I_k^H) y_{x,y}^m$$

$$\Delta Bias_k^H) = d(I_k^H)$$

【０１７９】
サブサンプル層Sのm番目の特徴マップの中の行x、列yの出力バイアスは、次の式によって決定される。

【０１８０】
【数３２】

$$d(O_{x,y}^{S,m}) = \sum_k^{170} d(I_{m,x,y}^H) W_{m,x,y}^{H,k}$$

【０１８１】
サブサンプル層Sのm番目の特徴マップの中の行x、列yの入力バイアスは、次の式によって決定される。

【０１８２】
【数３３】

$$d(I_{x,y}^{S,m}) = \varphi'(v_k) d(O_{x,y}^{S,m})$$

【０１８３】
サブサンプル層Sおよび畳み込み層Cのm番目の特徴マップの中の行x、列yの中の重みおよびバイアスのばらつきは、次の式によって決定される。

【０１８４】
【数３４】

$$\Delta W^{S,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{[x/2],[y/2]}^{S,m}) O_{x,y}^{C,m}$$

$$\Delta Bias^{S,m}) = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(O_{x,y}^{S,m})$$

【０１８５】
畳み込み層Cのk番目の特徴マップの中の行x、列yの出力バイアスは、次の式によって決定される。

【０１８６】
【数３５】

$$d(O_{x,y}^{C,k}) = d(I_{[x/2],[y/2]}^{S,k}) W^k$$

【０１８７】
畳み込み層Cのk番目の特徴マップの中の行x、列yの入力バイアスは、次の式によって決定される。

【０１８８】

【数３６】

$$d(I_{x,y}^{C,k}) = \varphi'(v_k)d(O_{x,y}^{C,k})$$

【０１８９】

l番目の畳み込み層Cのk番目の特徴マップのm番目の畳み込みコアの中の行r、列cにおける重みおよびバイアスのばらつき:

【０１９０】

【数３７】

$$\Delta W_{r,c}^{k,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k}) O_{x+r,y+c}^{l-1,m}$$

$$\Delta Bias^{C,k}) = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k})$$

【０１９１】

［残差接続］

図1Gは、特徴マップ追加を介して以前の情報ダウンストリームを再注入する残差接続を図示する。残差接続は、過去の出力テンソルをより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ダウンストリームフローへと再注入することを備え、このことは、データ処理フローに沿った情報の喪失を防ぐのを助ける。残差接続は、あらゆる大規模な深層学習モデルを悩ませる2つの一般的な問題、すなわち、勾配消失および表現上のボトルネック(representational bottleneck)に対処する。一般に、10層を超える層を有するあらゆるモデルに残差接続を追加することが有益である可能性が高い。上で論じられたように、残差接続は、より前の層の出力をより後の層への入力として利用可能にして、逐次ネットワークにおけるショートカットを実質的に作成することを備える。より前の出力は、より後の活性化に連結されるのではなく、より後の活性化と加算され、このことは両方の活性化が同じサイズであると想定している。それらが異なるサイズである場合、より前の活性化を目標の形状へと再成形するための線形変換が使用され得る。残差接続についての追加の情報は、本明細書に完全に記載されるかのようにすべての目的で参照によって本明細書において引用される、K.He、X.Zhang、S.Ren、およびJ.Sun、「DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION」、arXiv:1512.03385、2015において見出され得る。

【０１９２】

［残差学習およびスキップ接続］

図1Hは、残差ブロックおよびスキップ接続の一実装形態を示す。残差学習の主な考え方は、残差マッピングが元のマッピングよりはるかに簡単に学習されるということである。残差ネットワークは、訓練の正確さの劣化を軽減するために、いくつかの残差ユニットを積層する。残差ブロックは、深層ニューラルネットワークにおける勾配消失をなくすために、特別な追加のスキップ接続を利用する。残差ブロックの初めにおいて、データフローは2つのストリームへと分離され、第1のストリームがブロックの変更されない入力を搬送し、一方で第2のストリームが重みおよび非線形性を適用する。ブロックの終わりにおいて、2つのストリームは要素ごとの和を使用して統合される。そのような構築の主な利点は、勾配がより簡単にネットワークを通って流れることが可能になることである。残差ブロックおよびスキップ接続についての追加の情報は、A.V.D.Oord、S.Dieleman、H.Zen、K.Simonyan、O.Vinyals、A.Graves、N.Kalchbrenner、A.Senior、およびK.Kavukcuoglu、「WAVENET: A GENERATIVE MODEL FOR RAW AUDIO」、arXiv:1609.03499、2016において見出され得る。

【０１９３】
残差ネットワークから利益を得て、深層畳み込みニューラルネットワーク(CNN)を簡単に訓練することができ、画像分類および物体検出の精度改善を達成することができる。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(l+1)番目の層への入力として接続し、これは、以下の層遷移、すなわち$x_l=H_l(x_{l-1})$を生じさせる。残差ブロックは、恒等関数:$x_l=H_l(x_{l-1})+x_{l-1}$を用いて非線形変換をバイパスするスキップ接続を追加する。残差ブロックの利点は、勾配がより後の層からより前の層へ恒等関数を通って直接流れることができるということである。しかしながら、恒等関数およびH_lの出力は加算によって合成され、これはネットワークにおける情報フローを妨げることがある。

【０１９４】
［拡張畳み込み］
図10は拡張畳み込みを示す。膨張畳み込みとも呼ばれることのある拡張畳み込みは、字面上は「穴を伴う」を意味する。フランス語のalgorithme a trousが名称の由来であり、これは高速二項ウェーブレット変換を計算する。これらのタイプの畳み込み層では、フィルタの受容野に対応する入力は隣り合う点ではない。これが図10に示されている。入力間の距離は拡張係数に依存する。

【０１９５】
［WaveNet］
WaveNetは、生のオーディオ波形を生成するための深層ニューラルネットワークである。WaveNetは他の畳み込みネットワークから区別され、それは、WaveNetは低コストで比較的大きい「視覚野」を取り込むことが可能であるからである。その上、信号の条件をローカルおよびグローバルに追加することが可能であり、これにより、WaveNetが複数の声を伴うテキストツースピーチ(TTS)エンジンとして使用されることが可能になり、TTSはローカル条件および特定の声およびグローバル条件を与える。

【０１９６】
WaveNetの主なビルディングブロックは、因果的拡張畳み込みである。因果的拡張畳み込みの延長として、WaveNetは、図1Pに示されるようなこれらの畳み込みの積層を可能にする。この図において拡張畳み込みを用いて同じ受容野を取得するには、別の拡張層が必要である。積層は拡張畳み込みの反復であり、拡張畳み込み層の出力を単一の出力に接続する。これにより、WaveNetが比較的低い計算コストで1つの出力ノードの大きな「視覚」野を得ることが可能になる。比較のために、512個の入力の視覚野を得るには、完全畳み込みネットワーク(FCN)は511個の層を必要とする。拡張畳み込みネットワークの場合、8個の層が必要である。積層された拡張畳み込みは、2層の積層では7個の層、または4個の積層では6個の層しか必要ではない。同じ視覚野をカバーするために必要な計算能力の差の考え方を得るために、以下の表は、層当たり1つのフィルタおよび2というフィルタ幅という仮定のもとで、ネットワークにおいて必要とされる重みの数を示す。さらに、ネットワークが8ビットのバイナリ符号化を使用していることが仮定される。

【０１９７】
【表１】

ネットワークタイプ	積層数	チャネル当たりの重みの数	重みの総数
FCN	1	$2.6 . 10^5$	$2.6 . 10^6$
WN	1	1022	8176
WN	2	1022	8176
WN	4	508	4064

【０１９８】
WaveNetは、残差接続が行われる前にスキップ接続を追加し、これはすべての後続の残差ブロックをバイパスする。これらのスキップ接続の各々は、それらを一連の活性化関数および畳み込みに通す前に加算される。直観的には、これは各層において抽出される情報の合計である。
【０１９９】
［バッチ正規化］
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必須の部分にすることによって、深層ネットワーク訓練を加速するための方法である。バッチ正規化は、訓練の間に時間とともに平均および分散が変化しても、データを適応的に正規化することができる。バッチ正規化は、訓練の間に見られるデータのバッチごとの平均と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。バッチ正規化の主な影響は、残差接続とよく似て、勾配伝播を助けるので、深層ネットワークを可能にするということである。一部の超深層ネットワークは、複数のバッチ正規化層を含む場合にのみ訓練することができる。バッチ正規化についての追加の情報は、本明細書に完全に記載されるかのようにすべての目的で参照によって本明細書において引用される、S.IoffeおよびC.Szegedy、「BATCH NORMALIZATION: ACCELERATING DEEP NETWORK TRAINING BY REDUCING INTERNAL COVARIATE SHIFT」、arXiv:1502.03167、2015において見出され得る。
【０２００】
バッチ正規化は、全結合層または畳み込み層のように、モデルアーキテクチャへと挿入され得るさらに別の層として見ることができる。バッチ正規化層は通常、畳み込み層または密結合層の後で使用される。バッチ正規化層は、畳み込み層または密結合層の前でも使用され得る。両方の実装形態が、開示される技術によって使用されることが可能であり、図1Lにおいて示されている。バッチ正規化層は軸引数を取り込み、軸引数は正規化されるべき特徴軸を指定する。この引数はデフォルトでは1であり、これは入力テンソルにおける最後の軸である。これは、data_formatが「channels_last」に設定された状態でDense層、Conv1D層、RNN層、およびConv2D層を使用するときの正しい値である。しかし、data_formatが「channels_first」に設定されるConv2D層のニッチな使用事例では、特徴軸は軸1であり、バッチ正規化における軸引数は1に設定され得る。
【０２０１】
バッチ正規化は、入力をフィードフォワードすることと、バックワードパスを介してパラメータに関する勾配およびそれ自体の入力(its own input)を計算することとのための定義を提供する。実際には、バッチ正規化層は、畳み込み層または全結合層の後に挿入されるが、それは出力が活性化関数へと供給される前である。畳み込み層では、異なる位置にある同じ特徴マップの異なる要素、すなわち活性化が、畳み込みの性質に従うために同じ方法で正規化される。したがって、ミニバッチの中のすべての活性化は、活性化ごとにではなく、すべての位置にわたって正規化される。
【０２０２】
内部的な共変量のシフトは、深層アーキテクチャが訓練するのに時間がかかることで悪名高かった主な理由である。これは、深層ネットワークが各層において新しい表現を学習しなければならないだけではなく、それらの分布の変化も考慮しなければならないという事実によるものである。
【０２０３】
一般に共変量シフトは、深層学習の領域における既知の問題であり、現実世界の問題において頻繁に発生する。よくある共変量シフト問題は、最適ではない一般化性能につながり得る、訓練セットと検定セットでの分布の違いである。この問題は通常、標準化または白色化前処理ステップ(whitening preprocessing step)によって対処される。しかしながら、特に白色化動作は、計算負荷が高いので、共変量シフトが様々な層全体で発生する場合には特に、オンラインの状況では非現実的である。
【０２０４】

内部的な共変量シフトは、ネットワーク活性化の分布が、訓練の間のネットワークパラメータの変化により複数の層にわたって変化するという現象である。理想的には、各層は、各層が同じ分布を有するが機能的な関係は同じままであるような空間へと変換されるべきである。それぞれの層およびステップにおいてデータを脱相関および白色化するための、共分散行列の高価な計算を避けるために、各ミニバッチにわたる各層における各入力特徴量の分布を、平均が0になり標準偏差が1になるように正規化する。

【０２０５】

フォワードパス

フォワードパスの間、ミニバッチの平均および分散が計算される。これらのミニバッチの統計により、データは、平均を差し引き、標準偏差で除算することによって正規化される。最後に、データは、学習されたスケールおよびシフトパラメータを用いて、スケーリングおよびシフトされる。バッチ正規化フォワードパスf_{BN}が図1Iに図示されている。

【０２０６】

図1Iにおいて、それぞれ、μ_Bはバッチ平均であり、

【０２０７】

【数３８】

$$\sigma_\beta^2$$

【０２０８】

はバッチ分散である。学習されたスケールおよびシフトパラメータは、それぞれγおよびβと表記される。分かりやすくために、バッチ正規化手順は、本明細書では活性化ごとに説明され、対応するインデックスを省略する。

【０２０９】

正規化は微分可能な変換であるので、誤差はこれらの学習されたパラメータへと伝播され、したがって、恒等変換を学習することによってネットワークの再現能力を復元することが可能である。逆に、対応するバッチ統計と同一のスケールおよびシフトパラメータを学習することによって、それが実行すべき最適な操作であった場合、バッチ正規化変換はネットワークに対する効果を持たない。検定時に、バッチ平均および分散はそれぞれの母集団の統計により置き換えられ、それは、入力がミニバッチからの他のサンプルに依存しないからである。別の方法は、訓練の間にバッチ統計の平均をとり続け、検定時にこれらを使用してネットワーク出力を計算することである。検定時において、バッチ正規化変換は、図1Jに示されるように表現され得る。図1Jにおいて、μ_Dおよび

【０２１０】

【数３９】

$$\sigma_D^2$$

【０２１１】

は、バッチ統計ではなく、それぞれ母集団の平均および分散を示す。

【０２１２】

バックワードパス

正規化は微分可能な演算であるので、バックワードパスは図1Kに図示されるように計算され得る。

【０２１３】

［1D畳み込み］

1D畳み込みは、図1Mに示されるように、ローカルの1Dパッチまたはサブ配列を配列から抽出する。1D畳み込みは、入力配列の中の時間的パッチから各出力タイムステップを取得する。1D畳み込み層は、配列の中のローカルパターンを認識する。同じ入力変換がパッチごとに実行されるので、入力配列の中のある場所において学習されるパターンは、異なる

場所においてより後に認識されることが可能であり、このことは、1D畳み込み層変換を時間的変換に対して不変にする。たとえば、サイズ5の畳み込みウィンドウを使用して塩基の配列を処理する1D畳み込み層は、長さ5以下の塩基配列を学習することが可能であるべきであり、入力配列の中の任意の文脈において塩基のモチーフを認識することが可能であるべきである。したがって、塩基レベルの1D畳み込みは、塩基の形態について学習することが可能である。

【０２１４】

［グローバル平均プーリング］

図1Nは、グローバル平均プーリング(GAP)がどのように機能するかを示す。グローバル平均プーリングは、スコアリングのために最後の層の中の特徴量の空間的な平均をとることによって、分類のための全結合(FC)層を置換するために使用され得る。これは、訓練負荷を低減し、過剰適合の問題をバイパスする。グローバル平均プーリングは、モデルの前に構造的を適用し、これはあらかじめ定められた重みを伴う線形変換と等価である。グローバル平均プーリングは、パラメータの数を減らし、全結合層をなくす。全結合層は通常、最もパラメータと接続の多い層であり、グローバル平均プーリングは、同様の結果を達成するのにはるかに低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考え方は、スコアリングのために各々の最後の層の特徴マップからの平均値を信頼性係数として生成し、直接ソフトマックス層に供給することである。

【０２１５】

グローバル平均プーリングは、(1)グローバル平均プーリング層の中に余剰のパラメータがないので、グローバル平均プーリング層において過剰適合が避けられる、(2)グローバル平均プーリングの出力は特徴マップ全体の平均であるので、グローバル平均プーリングは空間的な変換に対してよりロバストになる、および(3)ネットワーク全体のすべてのパラメータの50%超を通常は占める、全結合層の中の大量のパラメータにより、それらをグローバル平均プーリング層で置き換えることで、モデルのサイズを大きく低減することができ、これが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なものにする、という3つの利点を有する。

【０２１６】

最後の層の中のより強い特徴がより高い平均値を有することが予想されるので、グローバル平均プーリングは理にかなっている。いくつかの実装形態では、グローバル平均プーリングは、分類スコアのための代理として使用され得る。グローバル平均プーリングのもとでの特徴マップは、信頼性マップとして解釈されることが可能であり、特徴マップとカテゴリとの間の対応付けを強制することができる。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抽象化されている場合、特に有効であり得る。しかしながら、グローバル平均プーリング自体は、マルチレベル特徴が部分モデルのようなグループへと合成されるべきである場合には十分ではなく、これは、グローバル平均プーリングの後に、単純な全結合層または他の分類器を追加することによって最良に実行される。

【０２１７】

［ゲノミクスにおける深層学習］

遺伝的変異は、多くの疾患の説明を助け得る。ヒトはそれぞれが固有の遺伝コードを持ち、個人のグループ内には多くの遺伝的バリアントがある。有害な遺伝的バリアントの大半は、自然選択によってゲノムから枯渇している。どの遺伝的変異が病原性または有害である可能性が高いかを特定することが重要である。このことは、研究者が、病原性である可能性が高い遺伝的バリアントに注目し、多くの疾患の診断および治療を加速させることを助けるであろう。

【０２１８】

バリアントの性質および機能的な影響(たとえば、病原性)をモデル化することは重要であるが、ゲノミクスの分野においては難しい仕事である。機能的ゲノムシーケンシング技術の急速な進化にもかかわらず、バリアントの機能的な結果の解釈には、細胞タイプに固有の転写制御システムの複雑さが原因で、大きな困難が立ちはだかっている。

【０２１９】
過去数十年にわたる生化学技術の進化は、これまでよりもはるかに低いコストでゲノムデータを高速に生成する、次世代シーケンシング(NGS)プラットフォームをもたらした。そのような圧倒的に大量のシーケンシングされたDNAは、アノテーションが困難なままである。教師あり機械学習アルゴリズムは通常、大量のラベリングされたデータが利用可能であるときには性能を発揮する。バイオインフォマティクスおよび多くの他のデータリッチな訓練法では、インスタンスをラベリングするプロセスが高価である。しかしながら、ラベリングされていないインスタンスは、安価であり容易に利用可能である。ラベリングされたデータの量が比較的少なく、ラベリングされていないデータの量がかなり多いシナリオでは、半教師あり学習が、手動のラベリングに対する費用対効果の高い代替手法となる。

【０２２０】
バリアント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ために、半教師ありアルゴリズムを使用する機会が生じる。人間の診断バイアスがない病原性バリアントのデータベースを得ることができる。

【０２２１】
病原性分類器に関して、深層ニューラルネットワークは、高水準の特徴を連続的にモデル化するために複数の非線形の複雑な変換層を使用する、あるタイプの人工ニューラルネットワークである。深層ニューラルネットワークは、観測される出力と予測される出力との差を搬送する逆伝播を介してフィードバックを提供し、パラメータを調整する。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、並列および分散コンピューティングの能力、および洗練された訓練アルゴリズムが利用可能になることとともに進化してきた。深層ニューラルネットワークは、コンピュータビジョン、音声認識、および自然言語処理などの、多数の領域において大きな進化を促進してきた。

【０２２２】
畳み込みニューラルネットワーク(CNN)および再帰型ニューラルネットワーク(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、およびプーリング層を備えるアーキテクチャにより、画像認識において特に成功してきた。再帰型ニューラルネットワークは、パーセプトロン、長短期メモリユニット、およびゲート付き回帰型ユニットのようなビルディングブロックの間で、巡回接続を用いて入力データの連続的情報を利用するように設計される。加えて、深層空間時間ニューラルネットワーク、多次元再帰型ニューラルネットワーク、および畳み込みオートエンコーダなどの、多くの他の新興の深層ニューラルネットワークが、限られた文脈に対して提案されている。

【０２２３】
深層ニューラルネットワークを訓練する目的は、各層における重みパラメータの最適化であり、このことは、最も適した階層的表現をデータから学習できるように、より単純な特徴を複雑な特徴へと徐々に合成する。最適化プロセスの単一のサイクルは次のように編成される。まず、ある訓練データセットのもとで、フォワードパスが各層の中の出力を順番に計算し、ネットワークを通じて関数信号を前に伝播させる。最後の出力層において、目的損失関数が、推論された出力と所与のラベルとの間の誤差を測定する。訓練誤差を最小にするために、バックワードパスは、連鎖律を逆伝播誤差信号に使用し、ニューラルネットワーク全体のすべての重みに関する勾配を計算する。最後に、重みパラメータは、確率的勾配降下に基づく最適化アルゴリズムを使用して更新される。一方、バッチ勾配降下は、各々の完全なデータセットに対するパラメータ更新を実行し、確率的勾配降下は、データ例の各々の小さいセットに対する更新を実行することによって確率的近似を提供する。いくつかの最適化アルゴリズムは、確率的勾配低下に由来する。たとえば、AdagradおよびAdam訓練アルゴリズムは、確率的勾配降下を実行しながら、それぞれ、各パラメータのための更新頻度および勾配のモーメントに基づいて学習率を適応的に修正する。

【０２２４】

深層ニューラルネットワークの訓練における別のコア要素は正則化であり、これは、過剰適応を避けることで良好な一般化性能を達成することを意図した戦略を指す。たとえば、重み減衰は、重みパラメータがより小さい絶対値へと収束するように、目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、訓練の間にニューラルネットワークから隠れユニットをランダムに除去し、可能性のあるサブネットワークのアンサンブルであると見なされ得る。ドロップアウトの能力を高めるために、新しい活性化関数であるmaxoutと、rnnDropと呼ばれる再帰型ニューラルネットワークのためのドロップアウトの変形が提案されている。さらに、バッチ正規化は、ミニバッチ内の各活性化のためのスカラー特徴量の正規化と、各平均および分散をパラメータとして学習することとを通じた、新しい正則化方法を提供する。

【０２２５】

シーケンシングされたデータが多次元かつ高次元であるとすると、深層ニューラルネットワークは、その広い適用可能性および高い予測能力により、バイオインフォマティクスの研究に対して高い将来性がある。畳み込みニューラルネットワークは、モチーフの発見、病原性バリアントの特定、および遺伝子発現の推論などの、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適合されてきた。畳み込みニューラルネットワークは、DNAを研究するのに特に有用である重み共有戦略を使用し、それは、この戦略が、重大な生物学的機能を有することが推定されるDNAにおける短い反復的なローカルパターンである配列モチーフを捉えることができるからである。畳み込みニューラルネットワークの特徴は、畳み込みフィルタの使用である。精巧に設計され人間により作られた特徴に基づく従来の分類手法とは異なり、畳み込みフィルタは、生の入力データを知識の有用な表現へとマッピングする処理と類似した、特徴の適応学習を実行する。この意味で、畳み込みフィルタは一連のモチーフスキャナとして機能し、それは、そのようなフィルタのセットが、入力の中の関連するパターンを認識し、訓練手順の間にそれらを更新することが可能であるからである。再帰型ニューラルネットワークは、タンパク質またはDNA配列などの、可変の長さの連続的データにおける長距離の依存関係を捉えることができる。

【０２２６】

したがって、バリアントの病原性を予測するための強力な計算モデルには、基礎科学研究と橋渡し研究の両方に対して莫大な利益があり得る。

【０２２７】

一般的な多型は、多世代の自然選択によりその健康性が試されてきた自然の実験結果を表している。ヒトのミスセンス置換と同義置換についてアレル頻度分布を比較すると、ヒト以外の霊長類の種における高いアレル頻度でのミスセンスバリアントの存在は、そのバリアントがヒトの集団においても自然選択を受けていることを高い信頼度で予測することを発見した。対照的に、より遠縁の種における一般的なバリアントは、進化的な距離が長くなるにつれて、負の選択を受ける。

【０２２８】

配列だけを使用して臨床的なde novoミスセンス変異を正確に分類する、半教師あり深層学習ネットワークを訓練するために、ヒト以外の6種の霊長類の種からの一般的な変異を利用する。500を超える既知の種により、霊長類の系統は、有意性が知られていない大半のヒトバリアントの影響を系統的にモデル化するのに、十分な一般的な変異を含んでいる。

【０２２９】

ヒト基準ゲノムには、7000万個のタンパク質を変化させる可能性のあるミスセンス置換が隠れており、それらの大半は、ヒトの健康への影響が特性把握されていない稀な変異である。これらの有意性が知られていないバリアントは、臨床上の応用においてゲノム解釈の課題となっており、集団全体にわたるスクリーニングおよび個別化医療のためのシーケンシングの長期的な採用の障害である。

【０２３０】

多様なヒトの集団にわたる一般的な変異の目録を作ることが、臨床的に良性の変異を特

定するのに有効な戦略であるが、現代のヒトから入手可能な一般的な変異は、我々の種の遠い過去におけるボトルネック事象により限られている。ヒトとチンパンジーは99%の配列相同性を共有しており、これは、チンパンジーバリアントに対して働く自然選択が、ヒトにおいて同一状態であるバリアントの影響をモデル化することの可能性を示唆している。ヒトの集団における自然な多型に対する平均合祖時間は、種の分岐時間の一部であるので、自然に発生するチンパンジー変異は大部分が、平衡選択により維持されるハプロタイプの稀な事例を除き、ヒト変異と重複しない変異空間に及ぶ。

【０２３１】

60706人のヒトからの集約されたエクソンデータが最近利用可能になったことで、ミスセンス変異と同義変異に対するアレル頻度スペクトラムを比較することによって、この仮説を検定することが可能になった。ExACにおけるシングルトンバリアントは、トリヌクレオチドコンテクストを使用して変異率を調整した後のde novo変異により予測される、予想される2.2:1のミスセンス:同義比とよく一致するが、より高いアレル頻度では、観察されるミスセンスバリアントの数は、自然選択による有害なバリアントの除去により減少する。アレル頻度スペクトラムにわたるミスセンス:同義比のパターンは、集団における頻度が0.1%未満であるミスセンスバリアントの大部分が軽度に有害である、すなわち、集団からの即刻の除去を保証するほど病原性が高くなく、高いアレル頻度で存在することが許容されるほど中立的でもないということを示しており、これはより限られた集団データに対する以前の観察と一致している。これらの発見は、0.1%～1%より高いアレル頻度を伴うバリアントを、平衡選択および創始者効果により引き起こされるよく記録されている少数の例外を除いて、浸透性の遺伝性疾患に対しては良性である可能性が高いものとして除去するという、診療室において広く行われている経験的な実践を支持するものである。

【０２３２】

この分析を、一般的なチンパンジーバリアント(チンパンジー集団のシーケンシングにおいて1回よりも多く観察される)と同一状態であるヒトバリアントのサブセットについて繰り返すと、ミスセンス:同義比は、アレル頻度スペクトラムにわたって概ね一定であることを発見した。チンパンジーの集団におけるこれらのバリアントの高いアレル頻度は、これらのバリアントがチンパンジーの自然選択のふるいにすでにかけられてきたことを示し、ヒトの集団における健康へのそれらのバリアントの中立的な影響は、ミスセンスバリアントに対する選択圧力が2つの種において高度に合致していることの注目すべき証拠を与えている。チンパンジーにおいて観察されるより低いミスセンス:同義比は、軽度に有害なバリアントの効率的な除去を可能にする先祖のチンパンジーの集団におけるより大きい実効集団サイズと一貫している。

【０２３３】

対照的に、稀なチンパンジーバリアント(チンパンジー集団のシーケンシングにおいて1回しか観察されない)は、より高いアレル頻度において、ミスセンス:同義比のあまり大きくない低下を示す。ヒト変異データからの同一サイズのコホートをシミュレートすると、このサイズのコホートにおいて一度観察されるバリアントの64%しか、集団全体において0.1%より高いアレル頻度を有せず、それと比べて、コホートにおいて複数回見られるバリアントについては99.8%が集団全体において0.1%より高いアレル頻度を有することが推定され、これは、稀なチンパンジーバリアントのすべてが選択のふるいにかけられたとは限らないことを示している。全体として、確認されたチンパンジーミスセンスバリアントの16%が、集団全体において0.1%未満のアレル頻度を有し、より高いアレル頻度では負の選択を受けることが推定される。

【０２３４】

次に、他のヒト以外の霊長類の種(ボノボ、ゴリラ、オランウータン、アカゲザル、およびマーモセット)において観察される変異と同一状態であるヒトバリアントを特徴付ける。チンパンジーと同様に、少数の稀なバリアント(約5～15%)の包含によるものであると推測される高いアレル頻度におけるミスセンス変異のわずかな枯渇を除き、ミスセンス:同義比がアレル頻度スペクトラムにわたって概ね等しいことを認めた。これらの結果は、

ミスセンスバリアントに対する選択圧が、ヒトの祖先の系統から約3500万年前に分岐したと推定される新世界ザルまでは少なくとも、霊長類の系統内で概ね合致していることを示唆する。

【０２３５】

他の霊長類におけるバリアントと同一状態であるヒトミスセンスバリアントは、ClinVarにおける良性の結果に対して強くエンリッチメントされる。未知のまたは矛盾するアノテーションを伴うバリアントを除いた後で、霊長類オーソログを伴うヒトバリアントは、ClinVarにおいて良性または良性の可能性が高いものとしてアノテートされる確率が約95%であり、それと比較して、ミスセンス変異全般では45%であることが観察される。ヒト以外の霊長類から病原性であるものとして分類されるClinVarバリアントの小さな割合は、健康なヒトの同様のサイズのコホートからの稀なバリアントを確認することにより観察されるであろう病原性のClinVarバリアントの割合と同程度である。大きなアレル頻度データベースの出現の前に分類を受けた、病原性であるまたは病原性である可能性が高いものとしてアノテートされたこれらのバリアントのかなりの割合が、今日では異なるように評価される可能性がある。

【０２３６】

ヒトの遺伝学の分野は、ヒト変異の臨床上の影響を推論するためにモデル生物に長い間依存してきたが、大半の遺伝的に扱いやすい動物モデルまでの進化的距離が長いことで、これらの発見がヒトに対してどの程度一般化可能であるかについての懸念が生まれている。ヒトおよびより遠縁の種におけるミスセンスバリアントに対する自然選択の合致を調査するために、4種の追加の哺乳類の種(ネズミ、ブタ、ヤギ、ウシ)と2種のより遠縁の脊椎動物(ニワトリ、ゼブラフィッシュ)からの概ね一般的な変異を含めるように、霊長類の系統を超えて分析を拡張した。以前の霊長類の分析とは対照的に、進化的距離が遠い場合には特に、稀なアレル頻度と比較して一般的なアレル頻度ではミスセンス変異が顕著に枯渇していることが観察され、これは、より遠縁の種における一般的なミスセンス変異のかなりの割合が、ヒトの集団においては負の選択を受けるであろうことを示している。それでも、より遠縁の脊椎動物におけるミスセンスバリアントの観察は、良性の結果の確率を高め、それは、自然選択により枯渇した一般的なミスセンスバリアントの割合は、基準であるヒトミスセンスバリアントに対して約50%よりはるかに低い枯渇率であるからである。これらの結果と一致して、ネズミ、イヌ、ブタ、およびウシにおいて観察されたヒトミスセンスバリアントは、ClinVarにおいて良性または良性の可能性が高いものとしてアノテートされる確率が約85%であり、それと比較して、霊長類の変異に対しては95%、ClinVarデータベース全体に対しては45%であることを発見した。

【０２３７】

様々な進化的距離にある近縁の種のペアの存在も、ヒトの集団における固定されたミスセンス置換の機能的な結果を評価するための機会を与える。哺乳類の系図上で近縁の種のペア(枝長<0.1)内で、固定されたミスセンス変異が、稀なアレル頻度と比較して一般的なアレル頻度で枯渇することが観察され、これは、複数の種にわたる固定された置換のかなりの割合が、霊長類の系統内であってもヒトにおいては非中立的であることを示している。ミスセンスの枯渇の程度の比較は、複数の種にわたる固定された置換が、同一種内の多型よりはるかに中立的ではないことを示している。興味深いことに、近縁の哺乳類間での複数の種にわたる変異は、同一種内の一般的な多型と比較して、ClinVarにおいてはさほどより病原性ではなく(良性または良性の可能性が高いものとしてアノテートされる確率が83%)、これらの変化がタンパク質の機能を無効にするのではなく、むしろ、種固有の適応的な利益を授けるタンパク質機能の調整を招いていることを示唆する。

【０２３８】

有意性が知られてない多数の潜在的なバリアントがあること、および臨床上の応用には正確なバリアント分類が決定的に重要であることにより、機械学習を用いた問題の解決が多く試みられてきたが、これらの努力は、一般的なヒトバリアントの量が不十分であること、および精選されたデータベースにおけるアノテーションの品質が疑わしいことにより

大きく制約されてきた。6種のヒト以外の霊長類からの変異は、一般的なヒト変異と重複せず大部分が良性の結果をもたらす300000個を超える固有のミスセンスバリアントに寄与し、機械学習手法に使用できる訓練データセットのサイズを大きく拡大した。

【０２３９】

人間により加工された多数の特徴およびメタ分類器を利用するこれまでのモデルと異なり、対象のバリアントの側にあるアミノ酸配列および他の種におけるオーソロガスな配列アラインメントのみを入力として取り込む、単純な深層学習残差ネットワークを適用する。タンパク質構造についての情報をネットワークに提供するために、配列だけから二次構造および溶媒接触性を学習するように2つの別々のネットワークを訓練し、これらをサブネットワークとしてより大きな深層学習ネットワークに組み込み、タンパク質構造に対する影響を予測する。配列を開始点として使用することで、不完全に確認されている可能性がある、または矛盾して適用されている可能性がある、タンパク質構造および機能ドメインのアノテーションにおける存在し得るバイアスが回避される。

【０２４０】

良性である可能性が高い霊長類バリアントと、変異率およびシーケンシングカバレッジについて一致するランダムな未知のバリアントとを分離するように、ネットワークのアンサンブルを最初に訓練することによって、訓練セットが良性のラベルを持つバリアントしか含まないという問題を克服するために、半教師あり学習を使用する。このネットワークのアンサンブルは、未知のバリアントの完全なセットをスコアリングするために、および、より病原性であるという予測される結果を持つ未知のバリアントに向かってバイアスをかけることによって分類器の次の反復をシードするように未知のバリアントの選択に影響を与えるために使用され、モデルが準最適な結果へと尚早に収束するのを防ぐために各反復において緩やかなステップをとる。

【０２４１】

一般的な霊長類の変異はまた、メタ分類器の増殖により客観的に評価することが難しくなっている既存の方法を評価するための、以前に使用された訓練データとは完全に無関係であるクリーンな評価データセットを提供する。10000個の提供された霊長類の一般的なバリアントを使用して、4つの他の人気のある分類アルゴリズム(Sift、Polyphen2、CADD、M-CAP)とともに、我々のモデルの性能を評価した。すべてのヒトミスセンスバリアントの概ね50%は、一般的なアレル頻度では自然選択によって除去されるので、変異率によって、10000個の提供された霊長類の一般的なバリアントと一致したランダムに選ばれたミスセンスバリアントのセットに対して、各分類器について50パーセンタイルのスコアを計算し、その閾値を使用して、提出された霊長類の一般的なバリアントを評価した。我々の深層学習モデルの正確さは、ヒトの一般的なバリアントだけで訓練された深層学習ネットワークを使用しても、またはヒトの一般的なバリアントと霊長類のバリアントの両方を使用しても、この独立の評価データセットについて、他の分類器よりはるかに良好であった。

【０２４２】

最近のトリオシーケンシング研究は、神経発達障害を持つ患者と患者の健康な兄弟における数千個のde novo変異の目録を作っており、症例群vs対照群におけるde novoミスセンス変異を分離する際の様々な分類アルゴリズムの強さの評価を可能にしている。4つの分類アルゴリズムの各々について、症例群vs対照群における各de novoミスセンスバリアントをスコアリングし、2つの分布の間の差のウィルコクソンの順位和検定からのp値を報告し、この臨床シナリオでは、霊長類バリアントについて訓練された深層学習方法(p約10^{-33})が他の分類器(p約10^{-13}から10^{-19})はるかに良好な性能であったことを示した。このコホートについて以前に報告された予想を超える、de novoミスセンスバリアントの約1.3-foldエンリッチメントから、およびミスセンスバリアントの約20%が機能喪失の影響を生むという以前の推定から、完璧な分類器はp約10^{-40}というp値で2つのクラスを分離することが予想され、これは我々の分類器に改善の余地がまだあることを示している。

【０２４３】

深度学習分類器の正確さは訓練データセットのサイズと符合し、6種の霊長類の各々からの変異データは独立に、分類器の正確さを上げることに寄与する。ヒト以外の霊長類の種が多数かつ多様にあることは、タンパク質を変化させるバリアントに対する選択圧力が霊長類の系統内で概ね合致していることを示す証拠とともに、臨床上のゲノム解釈を現在制約している、有意性が知られていない数百万個のヒトバリアントを分類するための効果的な戦略として、系統的な霊長類集団のシーケンシングを示唆する。504種の知られているヒト以外の霊長類の種のうち、約60%が狩猟および生息地喪失により絶滅に瀕しており、これらの固有の代わりのいない種と我々自身の両方に利益をもたらすであろう、緊急を要する世界的な保全の努力に対する動機となっている。

【０２４４】

ゲノムデータ全体はエクソンデータほど集約された形では利用可能ではないが、深いイントロン領域における自然選択の影響を検出するための能力を制限することで、エクソン領域から遠く離れた隠れたスプライシング変異の観察されるカウントと予想されるカウントを計算することも可能になった。全体として、エクソンイントロン境界から50ntを超える距離にある隠れたスプライシング変異において、60%の欠失を認めた。信号の減衰は、エクソンと比較してゲノムデータ全体ではサンプルサイズがより小さいことと、深いイントロンバリアントの影響を予測することがより難しいこととの組合せによるものである可能性が高い。

【０２４５】

［用語］

限定はされないが、特許、特許出願、論説、書籍、論文、およびウェブページを含む、本出願において引用されるすべての文献および同様の資料は、そのような文献および同様の資料のフォーマットとは無関係に、全体が参照によって明確に引用される。限定はされないが、定義される用語、用語の使用法、説明される技法などを含めて、引用される文献および同様の資料のうちの1つまたは複数が、本出願とは異なる場合、または本出願と矛盾する場合、本出願が優先する。

【０２４６】

本明細書では、以下の用語は示される意味を有する。

【０２４７】

塩基は、ヌクレオチド塩基またはヌクレオチド、すなわちA(アデニン)、C(シトシン)、T(チミン)、またはG(グアニン)を指す。

【０２４８】

本出願は、「タンパク質」および「翻訳配列」という用語を交換可能に使用する。

【０２４９】

本出願は、「コドン」および「塩基トリプレット」という用語を交換可能に使用する。

【０２５０】

本出願は、「アミノ酸」および「翻訳単位」という用語を交換可能に使用する。

【０２５１】

本出願は、「バリアント病原性分類器」、「バリアント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」、および「バリアント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」という語句を交換可能に使用する。

【０２５２】

「染色体」という用語は、生きている細胞の遺伝情報を持っている遺伝子の担体を指し、これはDNAおよびタンパク質の構成要素(特にヒストン)を備えるクロマチン鎖に由来する。従来の国際的に認識されている個々のヒトゲノム染色体ナンバリングシステムが本明細書で利用される。

【０２５３】

「サイト」という用語は、基準ゲノム上の一意な場所(たとえば、染色体ID、染色体の場所および向き)を指す。いくつかの実装形態では、サイトは、残基、配列タグ、または配列上のセグメントの場所であり得る。「座」という用語は、基準染色体上での核酸配列

または多型の具体的な位置を指すために使用され得る。

【０２５４】

本明細書の「サンプル」という用語は、典型的には、シーケンシングおよび/もしくはフェージングされるべき少なくとも1つの核酸配列を含有する核酸もしくは核酸の混合物を含有する、体液、細胞、組織、器官、または生物体に由来する、サンプルを指す。そのようなサンプルは、限定はされないが、唾液/口腔液、羊水、血液、血液の断片、細針生検サンプル(たとえば、直視下生検、細針生検など)、尿、腹膜液、胸膜液、組織外植、器官培養、および任意の他の組織もしくは細胞の標本、またはそれらの一部もしくはそれらの派生物、またはそれらから分離されたものを含む。サンプルはしばしば、ヒト対象(たとえば、患者)から取られるが、サンプルは、限定はされないが、イヌ、ネコ、ウマ、ヤギ、ヒツジ、ウシ、ブタなどを含む、染色体を有する任意の生物体から取ることができる。サンプルは、生物学的な供給源から得られるものとして直接使用されることがあり、または、サンプルの特性を修正するための前処理の後に使用されることがある。たとえば、そのような前処理は、血液から血漿を調製すること、粘液を希釈することなどを含み得る。前処理の方法はまた、限定はされないが、濾過、沈殿、希釈、蒸留、混合、遠心分離、凍結、凍結乾燥、濃縮、増幅、核酸断片化、干渉する要素の不活性化、試薬の追加、溶解などを伴い得る。

【０２５５】

「配列」という用語は、互いに結合されたヌクレオチドの鎖を含み、または表す。ヌクレオチドはDNAまたはRNAに基づき得る。1つの配列は複数の部分配列を含み得ることを理解されたい。たとえば、(たとえばPCRアンプリコン)の単一配列は350個のヌクレオチドを有し得る。サンプルリードは、これらの350個のヌクレオチド内の複数の部分配列を含み得る。たとえば、サンプルリードは、たとえば20～50個のヌクレオチドを有する、第1および第2のフランキング部分配列を含み得る。第1および第2のフランキング部分配列は、対応する部分配列(たとえば、40～100個のヌクレオチド)を有する反復的なセグメントの両側に位置し得る。フランキング部分配列の各々は、プライマー部分配列(たとえば、10～30個のヌクレオチド)を含み得る(またはその一部を含み得る)。読むのを簡単にするために、「部分配列」という用語は「配列」と呼ばれるが、2つの配列は必ずしも共通の鎖上で互いに別々であるとは限らないことを理解されたい。本明細書で説明される様々な配列を区別するために、配列は異なるラベル(たとえば、標的配列、プライマー配列、フランキング配列、基準配列など)を与えられ得る。「アレル」などの他の用語は、同様の物を区別するために異なるラベルを与えられ得る。

【０２５６】

「ペアエンドシーケンシング(paired-end sequencing)」という用語は、標的フラグメントの両端をシーケンシングするシーケンシング方法を指す。ペアエンドシーケンシングは、ゲノム再配置および反復セグメント、ならびに遺伝子融合および新規転写物の検出を容易にし得る。ペアエンドシーケンシングの方法論は、各々が本明細書において参照によって引用される、国際特許出願公開第WO07010252号、国際特許出願第PCTGB2007/003798号、および米国特許出願公開第2009/0088327号において説明されている。一例では、一連の操作は次のように実行され得る。(a)核酸のクラスタを生成する。(b)核酸を直線化する。(c)第1のシーケンシングプライマーをハイブリダイゼーションし、上で記載されたような延長、走査、およびデブロッキングの繰り返されるサイクルを実行する。(d)相補的なコピーを合成することによってフローセル表面上の標的核酸を「逆にする」。(e)再合成された鎖を直線化する。(f)第2のシーケンシングプライマーをハイブリダイゼーションし、上で記載されたような延長、走査、およびデブロッキングの繰り返されるサイクルを実行する。この逆転操作は、ブリッジ増幅の単一サイクルについて上に記載されたように試薬を導入するために実行され得る。

【０２５７】

「基準ゲノム」または「基準配列」という用語は、対象からの特定された配列の基準にするために使用され得る任意の生物体の任意の特定の既知のゲノム配列を、それが部分的

なものであるか完全なものであるかにかかわらず指す。たとえば、ヒト対象ならびに多くの他の生物体のために使用される基準ゲノムは、ncbi.nlm.nih.govの米国国立生物工学情報センターにおいて見つかる。「ゲノム」は、核酸配列で表現される、生物体またはウイルスの完全な遺伝情報を指す。ゲノムは、遺伝子とDNAのノンコーディング配列の両方を含む。基準配列は、それとアラインメントされるリードより大きいことがある。たとえば、それは少なくとも約100倍大きいことがあり、または少なくとも約1000倍大きいことがあり、または少なくとも約10000倍大きいことがあり、または少なくとも約105倍大きいことがあり、または少なくとも約106倍大きいことがあり、または少なくとも約107倍大きいことがある。一例では、基準ゲノム配列は、完全な長さのヒトゲノムの基準ゲノム配列である。別の例では、基準ゲノム配列は、13番染色体などの特定のヒト染色体に限定される。いくつかの実装形態では、基準染色体は、ヒトゲノムバージョンhg19からの染色体配列である。そのような配列は染色体基準配列と呼ばれ得るが、基準ゲノムという用語がそのような配列を包含することが意図される。基準配列の他の例には、他の種のゲノム、ならびに任意の種の染色体、部分染色体領域(鎖など)などがある。様々な実装形態において、基準ゲノムは、複数の個体に由来するコンセンサス配列または他の組合せである。しかしながら、いくつかの適用例では、基準配列は特定の個体から取られることがある。

【０２５８】

「リード」という用語は、ヌクレオチドサンプルまたは基準のフラグメントを記述する配列データの集合体を指す。「リード」という用語は、サンプルリードおよび/または基準リードを指し得る。通常、必須ではないが、リードは、サンプルまたは基準における連続的な塩基対の短い配列を表す。リードは、サンプルまたは基準フラグメントの塩基対配列によって文字で(ATCGで)表され得る。リードは、メモリデバイスに記憶され、リードが基準配列と一致するかどうか、または他の基準を満たすかどうかを決定するために適宜処理され得る。リードは、シーケンシング装置から直接、またはサンプルに関する記憶された配列情報から間接的に得られ得る。いくつかの場合、リードは、たとえば染色体またはゲノム領域または遺伝子にアラインメントされ具体的に割り当てられ得る、より大きい配列または領域を特定するために使用され得る、十分な長さの(たとえば、少なくとも約25bp)DNA配列である。

【０２５９】

次世代シーケンシング方法には、たとえば、合成技術によるシーケンシング(Illumina)、パイロシーケンシング(454)、イオン半導体技術(Ion Torrentシーケンシング)、単一分子リアルタイムシーケンシング(Pacific Biosciences)、およびライゲーションによるシーケンシング(SOLiDシーケンシング)がある。シーケンシング方法に応じて、各リードの長さは、約30bpから10000bp以上にまで変動し得る。たとえば、SOLiDシーケンサを使用するIlluminaシーケンシング方法は、約50bpの核酸リードを生成する。別の例では、Ion Torrentシーケンシングは最高で400bpの核酸リードを生成し、454パイロシーケンシングは約700bpの核酸リードを生成し得る。さらに別の例では、単一分子リアルタイムシーケンシング方法は、10000bpから15000bpのリードを生成し得る。したがって、いくつかの実装形態では、核酸配列リードは、30～100bp、50～200bp、または50～400bpの長さを有する。

【０２６０】

「サンプルリード」、「サンプル配列」、または「サンプルフラグメント」という用語は、サンプルからの対象のゲノム配列の配列データを指す。たとえば、サンプルリードは、フォワードプライマー配列およびリバースプライマー配列を有するPCRアンプリコンからの配列データを備える。配列データは、任意の配列選択方法から得られ得る。サンプルリードは、たとえば、sequencing-by-synthesis(SBS)反応、sequencing-by-ligation反応、または、そのために反復要素の長さおよび/または正体を決定することが望まれる任意の他の適切なシーケンシング方法からのものであり得る。サンプルリードは、複数のサンプルリードに由来するコンセンサス(たとえば、平均または加重)配列であり得る。いくつかの実装形態では、基準配列を提供することは、PCRアンプリコンのプライマー配列に基

づいて対象座を特定することを備える。

【０２６１】

「生フラグメント」という用語は、サンプルリードまたはサンプルフラグメント内で指定場所または二次的な対象場所と少なくとも部分的に重複する、対象のゲノム配列の部分に対する配列データを指す。生フラグメントの非限定的な例には、duplex stitchedフラグメント、simplex stitchedフラグメント、duplex un-stitchedフラグメント、およびsimplex un-stitchedフラグメントがある。「生」という用語は、生フラグメントがサンプルリードの中の潜在的なバリアントに対応しそれが本物であることを証明または確認する、支持バリアントを呈するかどうかにかかわらず、サンプルリードの中の配列データに対する何らかの関連を有する配列データを含むことを示すために使用される。「生フラグメント」という用語は、フラグメントが、サンプルリードの中のバリアントコールを妥当性確認する支持バリアントを必ず含むことを示さない。たとえば、サンプルリードが第1のバリアントを呈することが、バリアントコールアプリケーションによって決定されるとき、バリアントコールアプリケーションは、1つまたは複数の生フラグメントが、サンプルリードの中にそのバリアントがあるとすれば存在することが予想され得る対応するタイプの「支持」バリアントを欠いていることを決定し得る。

【０２６２】

「マッピング」、「アラインメントされる」、「アラインメント」、または「アラインメントしている」という用語は、リードまたはタグを基準配列と比較し、それにより、基準配列がリード配列を含むかどうかを決定するプロセスを指す。基準配列がリードを含む場合、リードは、基準配列にマッピングされることがあり、またはいくつかの実装形態では、基準配列の中の特定の位置にマッピングされることがある。いくつかの場合、アラインメントは単に、リードが特定の基準配列のメンバーであるかどうか(すなわち、リードが基準配列の中に存在するかしないか)を伝える。たとえば、ヒト13番染色体の基準配列に対するリードのアラインメントは、リードが13番染色体の基準配列の中に存在するかどうかを伝える。この情報を提供するツールは、セットメンバーシップテスターと呼ばれ得る。いくつかの場合、アラインメントは追加で、リードまたはタグがマッピングする基準配列の中の位置を示す。たとえば、基準配列がヒトゲノム配列全体である場合、アラインメントは、リードが13番染色体上に存在することを示すことがあり、さらに、リードが13番染色体の特定の鎖および/またはサイトにあることを示すことがある。

【０２６３】

「インデル」という用語は、生物体のDNAにおける塩基の挿入および/または欠失を指す。マイクロインデルは、1～50個のヌクレオチドの正味の変化をもたらすインデルを表す。ゲノムのコーディング領域において、インデルの長さが3の倍数ではない限り、インデルはフレームシフト変異を生み出す。インデルは点変異と対比され得る。インデルは配列からヌクレオチドを挿入または削除するが、点変異はDNAの全体の数を変えることなくヌクレオチドのうちの1つを置き換えるある形式の置換である。インデルは、タンデム塩基変異(TBM)とも対比することができ、TBMは隣接するヌクレオチドにおける置換として定義され得る(主に2つの隣接するヌクレオチドにおける置換、しかし3つの隣接するヌクレオチドにおける置換が観察されている)。

【０２６４】

「バリアント」という用語は、核酸基準と異なる核酸配列を指す。典型的な核酸配列バリアントには、限定はされないが、一塩基多型(SNP)、短い欠失および挿入の多型(インデル)、コピー数変異(CNV)、マイクロサテライトマーカー、またはショートタンデムリピートおよび構造変異がある。体細胞バリアントコーリング(somatic variant calling)は、DNAサンプルにおいて低頻度に存在するバリアントを特定するための試みである。体細胞バリアントコーリングは、癌治療の文脈において関心の対象である。癌はDNAの変異の蓄積により引き起こされる。腫瘍からのDNAサンプルは一般に異質であり、いくつかの正常細胞、癌進行の早期段階にあるいくつかの細胞(少数の変異を伴う)、およびいくつかの後期段階の細胞(多数の変異を伴う)を含む。この異質さにより、(たとえば、FFPEサンプルか

ら)腫瘍をシーケンシングするとき、体細胞突然変異がしばしば低頻度で現れる。たとえば、ある所与の塩基を含むリードの10%だけにおいて、SNVが見られることがある。バリアント分類器によって体細胞性または生殖細胞性であると分類されるべきバリアントは、「検定対象バリアント(variant under test)」とも本明細書では呼ばれる。

【０２６５】

「ノイズ」という用語は、シーケンシングプロセスおよび/またはバリアントコールアプリケーションにおける1つまたは複数のエラーに起因する誤ったバリアントコールを指す。

【０２６６】

「バリアント頻度」という用語は、割合または百分率で表される、ある集団の中の特定の座におけるアレル(遺伝子のバリアント)の相対的な頻度を表す。たとえば、この割合または百分率は、そのアレルを持つ集団の中のすべての染色体の割合であり得る。例として、サンプルバリアント頻度は、ある個人からの対象のゲノム配列について取得されたリードおよび/またはサンプルの数に対応する「集団」にわたる、対象のゲノム配列に沿った特定の座/場所におけるアレル/バリアントの相対的な頻度を表す。別の例として、基準バリアント頻度は、1つまたは複数の基準ゲノム配列に沿った特定の座/場所におけるアレル/バリアントの相対的な頻度を表し、リードおよび/またはサンプルの数に対応する「集団」は、正常な個人の集団からの1つまたは複数の基準ゲノム配列について取得される。

【０２６７】

「バリアントアレイ頻度(VAF)」という用語は、標的場所における、バリアントと一致することが観察されたシーケンシングされたリードをカバレッジ全体で割った百分率を指す。VAFはバリアントを持つシーケンシングされたリードの比率の尺度である。

【０２６８】

「場所」、「指定場所」、および「座」という用語は、ヌクレオチドの配列内の1つまたは複数のヌクレオチドの位置または座標を指す。「場所」、「指定場所」、および「座」という用語は、ヌクレオチドの配列の中の1つまたは複数の塩基対の位置または座標も指す。

【０２６９】

「ハプロタイプ」という用語は、一緒に受け継がれる染色体上の隣接するサイトにおけるアレルの組合せを指す。ハプロタイプは、所与の座のセット間で組み換え事象が発生した場合にはその数に依存して、1つの座、いくつかの座、または染色体全体であり得る。

【０２７０】

本明細書の「閾値」という用語は、サンプル、核酸、またはその一部(たとえば、リード)を特徴付けるためにカットオフとして使用される、数値または数字ではない値を指す。閾値は経験的な分析に基づいて変動し得る。閾値は、そのような値の示唆をもたらす源がある特定の方式で分類されるべきであるかどうかを決定するために、測定された値または計算された値と比較され得る。閾値は経験的または分析的に特定され得る。閾値の選択は、ユーザが分類を行うために有することを望む信頼性のレベルに依存する。閾値は特定の目的で(たとえば、感度と選択度のバランスをとるように)選ばれ得る。本明細書では、「閾値」という用語は、分析のコースが変更され得る点、および/または活動が惹起され得る点を示す。閾値は所定の数である必要はない。代わりに、閾値は、たとえば、複数の要因に基づく関数であり得る。閾値は状況に適応するものであり得る。その上、閾値は、上限、下限、または制限値間の範囲を示し得る。

【０２７１】

いくつかの実装形態では、シーケンシングデータに基づく尺度またはスコアが閾値と比較され得る。本明細書では、「尺度」または「スコア」という用語は、シーケンシングデータから決定された値もしくは結果を含むことがあり、または、シーケンシングデータから決定された値もしくは結果に基づく関数を含むことがある。閾値と同様に、尺度またはスコアは状況に適応するものであり得る。たとえば、尺度またはスコアは正規化された値であり得る。スコアまたは尺度の例として、1つまたは複数の実装形態は、データを分析

するときにカウントスコアを使用し得る。カウントスコアはサンプルリードの数に基づき得る。サンプルリードは1つまたは複数のフィルタリング段階を経ていることがあるので、サンプルリードは少なくとも1つの一般的な特性または品質を有する。たとえば、カウントスコアを決定するために使用されるサンプルリードの各々は、基準配列とアラインメントされていることがあり、または潜在的なアレルとして割り当てられることがある。一般的な特性を有するサンプルリードの数はリードカウントを決定するためにカウントされ得る。カウントスコアはリードカウントに基づき得る。いくつかの実装形態では、カウントスコアはリードカウントに等しい値であり得る。他の実装形態では、カウントスコアはリードカウントおよび他の情報に基づき得る。たとえば、カウントスコアは、遺伝子座の特定のアレルに対するリードカウントおよび遺伝子座に対するリードの総数に基づき得る。いくつかの実装形態では、カウントスコアは、遺伝子座に対するリードカウントおよび以前に得られたデータに基づき得る。いくつかの実装形態では、カウントスコアは複数の所定の値の間の正規化されたスコアであり得る。カウントスコアはまた、サンプルの他の座からのリードカウントの関数、または対象サンプルと同時に実行された他のサンプルからのリードカウントの関数であり得る。たとえば、カウントスコアは、特定のアレルのリードカウントおよびサンプルの中の他の座のリードカウントおよび/または他のサンプルからのリードカウントの関数であり得る。一例として、他の座からのリードカウントおよび/または他のサンプルからのリードカウントが、特定のアレルに対するカウントスコアを正規化するために使用され得る。

【０２７２】

「カバレッジ」または「フラグメントカバレッジ」という用語は、配列の同じフラグメントに対するサンプルリードの数のカウントまたは他の尺度を指す。リードカウントは対応するフラグメントをカバーするリードの数のカウントを表し得る。あるいは、カバレッジは、履歴の知識、サンプルの知識、座の知識などに基づく指定された係数を、リードカウントと乗じることによって決定され得る。

【０２７３】

「リード深さ」(慣習的に「×」が後に続く数)という用語は、標的場所における重複するアラインメントを伴うシーケンシングされたリードの数を指す。これはしばしば、平均として、または間隔(エクソン、遺伝子、またはパネルなど)のセットにわたってカットオフを超える百分率として表される。たとえば、パネル平均カバレッジが1.105×であり、カバーされる標的塩基の98%が>100×であるということを、臨床報告が述べることがある。

【０２７４】

「塩基コール品質スコア」または「Qスコア」という用語は、単一のシーケンシングされた塩基が正しい確率に反比例する、0～20の範囲のPHREDスケーリングされた確率を指す。たとえば、Qが20であるT塩基コールは、0.01という信頼性P値を伴い正しい可能性が高いと見なされる。Q<20であるあらゆる塩基コールは低品質であると見なされるべきであり、バリアントを支持するシーケンシングされたリードのかなりの部分が低品質であるようなあらゆる特定されたバリアントは、偽陽性の可能性があると見なされるべきである。

【０２７５】

「バリアントリード」または「バリアントリード数」という用語は、バリアントの存在を支持するシーケンシングされたリードの数を指す。

【０２７６】

［シーケンシングプロセス］

本明細書に記載される実装形態は、配列の変異を特定するために核酸配列を分析することに適用可能であり得る。実装形態は、遺伝子の場所/座の潜在的なバリアント/アレルを分析し、遺伝子座の遺伝子型を決定するために、言い換えると、座に対する遺伝子型コールを提供するために使用され得る。例として、核酸配列は、米国特許出願公開第2016/0085910号および米国特許出願公開第2013/0296175号において説明される方法およびシステムに従って分析されることがあり、これらの出願公開の完全な主題の全体が、本明細書にお

いて参照によって明確に引用される。

【０２７７】

一実装形態では、シーケンシングプロセスは、DNAなどの核酸を含む、または含むことが疑われるサンプルを受け取ることを含む。サンプルは、動物(たとえばヒト)、植物、バクテリア、または菌類などの、既知のまたは未知の源からのものであり得る。サンプルは源から直接採取され得る。たとえば、血液または唾液が個体から直接採取され得る。代わりに、サンプルは源から直接採取されないことがある。次いで、1つまたは複数のプロセッサは、シーケンシングのためのサンプルを調製するようにシステムに指示する。この調製は、外来の物質を除去することおよび/または何らかの物質(たとえば、DNA)を隔離することを含み得る。生体サンプルは、特定のアッセイのための特徴を含むように調製され得る。たとえば、生体サンプルは、sequencing-by-synthesis(SBS)のために調製され得る。いくつかの実装形態では、調製することは、ゲノムのいくつかの領域の増幅を含み得る。たとえば、調製することは、STRおよび/またはSNRを含むことが知られている所定の遺伝子座を増幅することを含み得る。遺伝子座は、所定のプライマー配列を使用して増幅され得る。

【０２７８】

次に、1つまたは複数のプロセッサは、サンプルをシーケンシングするようにシステムに指示する。シーケンシングは、様々な既知のシーケンシングプロトコルを通じて実行され得る。特定の実装形態では、シーケンシングはSBSを含む。SBSでは、複数の蛍光ラベリングされたヌクレオチドが、光学基板の表面(たとえば、フローセルの中のチャネルを少なくとも部分的に画定する表面)上に存在する増幅されたDNAの複数のクラスタ(場合によっては数百万個のクラスタ)をシーケンシングするために使用される。フローセルはシーケンシングのための核酸サンプルを含むことがあり、ここでフローセルは適切なフローセルホルダ内に配置される。

【０２７９】

核酸は、未知の標的配列に隣接する既知のプライマー配列を備えるように調製され得る。最初のSBSシーケンシングサイクルを開始するために、1つまたは複数の異なるようにラベリングされたヌクレオチド、およびDNAポリメラーゼなどが、流体サブシステムによってフローセルの中へと/フローセルを通って流され得る。単一のタイプのヌクレオチドが一度に追加されるか、または、シーケンシング手順において使用されるヌクレオチドが反転可能な末端の性質を持つように特別に設計されるかのいずれかであってよく、これにより、シーケンシング反応の各サイクルが、いくつかのタイプのラベリングされたヌクレオチド(たとえば、A、C、T、G)の存在下で同時に発生することが可能になる。ヌクレオチドは、蛍光色素などの検出可能なラベル部分を含み得る。4つのヌクレオチドが一緒に混合される場合、ポリメラーゼは組み込むべき正しい塩基を選択することが可能であり、各配列は一塩基だけ延長される。組み込まれないヌクレオチドは、洗浄液をフローセルに流すことによって洗い落とされ得る。1つまたは複数のレーザーが、核酸を励起して蛍光を誘導し得る。核酸から放出される蛍光は組み込まれた塩基の蛍光色素に基づき、異なる蛍光色素は異なる波長の放出光を放出し得る。デブロッキング試薬が、延長され検出されたDNA鎖から反転可能な末端グループを除去するためにフローセルに追加され得る。次いでデブロッキング試薬が、洗浄液をフローセルに流すことによって洗い落とされ得る。そうすると、フローセルは、上に記載されたようなラベリングされたヌクレオチドの導入で開始するシーケンシングのさらなるサイクルの準備ができる。流体および検出の操作は、シーケンシングの実行を完了するために何回か繰り返され得る。例示的なシーケンシング方法は、たとえば、Bentley他、Nature 456:53-59(2008)、国際特許出願公開第WO 04/018497号、米国特許第7057026号、国際特許出願公開第WO 91/06678号、国際特許出願公開第WO 07/123744号、米国特許第7329492号、米国特許第7211414号、米国特許第7315019号、米国特許第7405281号、および米国特許出願公開第2008/0108082号において説明されており、これらの各々が参照によって本明細書において引用される。

【０２８０】

いくつかの実装形態では、核酸は表面に付着され、シーケンシングの前または間に増幅され得る。たとえば、増幅は、表面上に核酸クラスタを形成するためにブリッジ増幅を使用して行われ得る。有用なブリッジ増幅方法は、たとえば、米国特許第5641658号、米国特許出願公開第2002/0055100号、米国特許第7115400号、米国特許出願公開第2004/0096853号、米国特許出願公開第2004/0002090号、米国特許出願公開第2007/0128624号、および米国特許出願公開第2008/0009420号において説明されており、これらの各々の全体が参照によって本明細書において引用される。表面上で核酸を増幅するための別の有用な方法は、たとえば、Lizardi他、Nat. Genet. 19:225-232(1998)、および米国特許出願公開第2007/0099208A1号において説明されるようなローリングサークル増幅(RCA)であり、これらの各々が参照によって本明細書において引用される。

【０２８１】

1つの例示的なSBSプロトコルは、たとえば、国際特許出願公開第WO 04/018497号、米国特許出願公開第2007/0166705A1号、および米国特許第7057026号において説明されるような、除去可能な3'ブロックを有する修正されたヌクレオチドを利用し、これらの各々が参照によって本明細書において引用される。たとえば、SBS試薬の反復されるサイクルが、たとえばブリッジ増幅プロトコルの結果として、標的核酸が付着されたフローセルに導入され得る。核酸クラスタは、直線化溶液を使用して単鎖の形態へと変換され得る。直線化溶液は、たとえば、各クラスタの1本の鎖を開裂することが可能な制限エンドヌクレアーゼを含み得る。とりわけ化学開裂(たとえば、過ヨード酸塩を用いたジオール結合の開裂)、熱またはアルカリへの曝露によるエンドヌクレアーゼ(たとえば、米国マサチューセッツ州イプスウィッチのNEBにより供給されるような「USER」、部品番号M5505S)を用いた開裂による無塩基サイトの開裂、そうされなければデオキシリボヌクレオチドからなる増幅産物へと組み込まれるリボヌクレオチドの開裂、光化学開裂またはペプチドリンカーの開裂を含む、開裂の他の方法が、制限酵素またはニッキング酵素に対する代替として使用され得る。直線化操作の後で、シーケンシングプライマーは、シーケンシングされるべき標的核酸へのシーケンシングプライマーのハイブリダイゼーションのための条件下で、フローセルに導入され得る。

【０２８２】

次いで、フローセルが、単一のヌクレオチドの追加によって各標的核酸にハイブリダイゼーションされるプライマーを延長するための条件下で、除去可能な3'ブロックおよび蛍光ラベルを伴う修正されたヌクレオチドを有するSBS延長試薬と接触させられ得る。単一のヌクレオチドだけが各プライマーに追加され、それは、修正されたヌクレオチドが、シーケンシングされているテンプレートの領域と相補的な成長中のポリヌクレオチド鎖へと組み込まれると、さらなる配列延長を指示するために利用可能な自由な3'-OH基がないので、ポリメラーゼがさらなるヌクレオチドを追加できないからである。SBS延長試薬は、除去され、放射線による励起のもとでサンプルを保護する構成要素を含む走査試薬により置き換えられ得る。走査試薬の例示的な構成要素は、米国特許出願公開第2008/0280773A1号および米国特許出願第13/018255号において説明され、これらの各々が参照によって本明細書に引用される。次いで、延長された核酸が、走査試薬の存在下で蛍光により検出され得る。蛍光が検出されると、3'ブロックが、使用されるブロッキンググループに適切なデブロック試薬を使用して除去され得る。それぞれのブロッキンググループに対して有用な例示的なデブロック試薬は、国際特許出願公開第WO004018497号、米国特許出願公開第2007/0166705A1号、および米国特許第7057026号において説明されており、これらの各々が参照によって本明細書において引用される。デブロック試薬は、3'OH基を有する延長されたプライマーにハイブリダイゼーションされる標的核酸を残して洗浄されてよく、このプライマーはこれで、さらなるヌクレオチドの追加が可能になる。したがって、延長試薬、走査試薬、およびデブロック試薬を追加するサイクルは、操作のうちの1つまたは複数の間の任意選択の洗浄とともに、所望の配列が得られるまで繰り返され得る。上記のサイクルは、修正されたヌクレオチドの各々に異なるラベルが付けられているとき、特定の塩基に対応することが知られている、サイクルごとに単一の延長試薬導入操作を使用して行わ

れ得る。異なるラベルが、各組み込み操作の間に追加されるヌクレオチドの区別を容易にする。代わりに、各サイクルは、延長試薬導入の別個の操作と、それに続く走査試薬導入と検出の別
個の操作とを含むことがあり、この場合、ヌクレオチドのうちの2つ以上が同じラベルを有することが可能であり、それらを導入の既知の順序に基づいて区別することができる。

【０２８３】

シーケンシング操作は特定のSBSプロトコルに関して上で論じられたが、シーケンシングのための他のプロトコルおよび様々な他の分子分析法のいずれもが、必要に応じて行われ得ることが理解されるであろう。

【０２８４】

次いで、システムの1つまたは複数のプロセッサは、後続の分析のためのシーケンシングデータを受け取る。シーケンシングデータは、.BAMファイルなどの様々な方式でフォーマットされ得る。シーケンシングデータは、たとえばいくつかのサンプルリードを含み得る。シーケンシングデータは、ヌクレオチドの対応するサンプル配列を有する複数のサンプルリードを含み得る。1つだけのサンプルリードが論じられるが、シーケンシングデータは、たとえば、数百個、数千個、数十万個、または数百万個のサンプルリードを含み得ることを理解されたい。異なるサンプルリードは異なる数のヌクレオチドを有し得る。たとえば、サンプルリードは、10個のヌクレオチドから約500個以上のヌクレオチドにまでわたり得る。サンプルリードは源のゲノム全体にわたり得る。一例として、サンプルリードは、疑わしいSTRまたは疑わしいSNPを有する遺伝子座などの、所定の遺伝子座の方を向いている。

【０２８５】

各サンプルリードは、サンプル配列、サンプルフラグメント、または標的配列と呼ばれ得る、ヌクレオチドの配列を含み得る。サンプル配列は、たとえば、プライマー配列、フランキング配列、および標的配列を含み得る。サンプル配列内のヌクレオチドの数は、30個、40個、50個、60個、70個、80個、90個、100個以上を含み得る。いくつかの実装形態では、サンプルリード(またはサンプル配列)のうちの1つまたは複数は、少なくとも150個のヌクレオチド、200個のヌクレオチド、300個のヌクレオチド、400個のヌクレオチド、500個のヌクレオチド、またはそれより多くを含む。いくつかの実装形態では、サンプルリードは、1000個を超えるヌクレオチド、2000個を超えるヌクレオチド、またはそれより多くを含み得る。サンプルリード(またはサンプル配列)は、一端または両端にプライマー配列を含み得る。

【０２８６】

次に、1つまたは複数のプロセッサは、シーケンシングデータを分析して、潜在的なバリアントコールおよびサンプルバリアントコールのサンプルバリアント頻度を取得する。この操作は、バリアントコールアプリケーションまたはバリアントコーラとも呼ばれ得る。したがって、バリアントコーラはバリアントを特定または検出し、バリアント分類器は検出されたバリアントを体細胞性または生殖細胞性であるものとして分類する。代替的なバリアントコーラが本明細書の実装形態に従って利用されることがあり、ここで、異なるバリアントコーラは、実行されているシーケンシング操作のタイプ、対象のサンプルの特徴などに基づき使用され得る。バリアントコールアプリケーションの1つの非限定的な例は、https://github.com/Illumina/Pisces においてホストされ、論説Dunn, TamsenおよびBerry, GwennおよびEmig-Agius, DorotheaおよびJiang, YuおよびIyer, AnitaおよびUdar, NitinおよびStromberg, Michael、(2017)、Pisces: An Accurate and Versatile Single Sample Somatic and Germline Variant Caller、595-595、10.1145/3107411.3108203において説明される、Illumina Inc.(カリフォルニア州サンディエゴ)によるPisces(商標)アプリケーションであり、上記の論説の完全な主題の全体が、参照によって本明細書において引用される。

【０２８７】

そのようなバリアントコールアプリケーションは4つの順番に実行されるモジュールを

備え得る。

【０２８８】

(1)Pisces Read Stithcer:BAMの中の対になっているリード(同じ分子のリード1およびリード2)をコンセンサスリードへとステッチングすることによってノイズを減らす。出力はステッチングされたBAMである。

【０２８９】

(2)Pisces Variant Caller:小さいSNV、挿入および欠失をコールする。Piscesは、リード境界、基本的なフィルタリングアルゴリズム、および単純なポワソンベースのバリアント信頼性スコアリングアルゴリズムによって分解される、合祖バリアントへのバリアント折り畳み(variant-collapsing)アルゴリズムを含む。出力はVCFである。

【０２９０】

(3)Pisces Variant Quality Recalibrator(VQR):バリアントコールが熱損傷またはFFPE脱アミノ化と関連付けられるパターンに圧倒的に従う場合、VQRステップは疑わしいバリアントコールのバリアントQスコアを下げる。出力は調整されたVCFである。

【０２９１】

(4)Pisces Variant Phaser(Scylla):クローンの亜集団から少数のバリアントを複雑なアレルへと組み立てるために、read-backed greedy clustering法を使用する。このことは、下流のツールによる機能的な結果のより正確な決定を可能にする。出力は調整されたVCFである。

【０２９２】

加えて、または代わりに、この操作は、https://github.com/Illumina/strelkaにおいてホストされ、論説T Saunders, ChristopherおよびWong, WendyおよびSwamy, SajaniおよびBecq, JenniferおよびJ Murray, LisaおよびCheetham, Keira、(2012)、Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs、Bioinformatics (Oxford, England)、28、1811-7、10.1093/bioinformatics/bts271において説明される、Illumina Inc.によるバリアントコールアプリケーションStrelka(商標)アプリケーションを利用することがあり、上記の論説の完全な主題の全体が、参照によって本明細書において明確に引用される。さらに、加えて、または代わりに、この操作は、https://github.com/Illumina/strelkaにおいてホストされ、論説Kim, S、Scheffler, K、Halpern, A.L.、Bekritsky, M.A、Noh, E、Kallberg, M、Chen, X、Beyter, D、Krusche, P、およびSaunders, C.T、(2017)、Strelka2: Fast and accurate variant calling for clinical sequencing applicationsにおいて説明される、Illumina Inc.によるバリアントコールアプリケーションStrelka2(商標)アプリケーションを利用することがあり、上記の文書の完全な主題の全体が、参照によって本明細書において明確に引用される。その上、加えて、または代わりに、この操作は、https://github.com/Illumina/Nirvana/wikiにおいてホストされ、論説Stromberg, MichaelおよびRoy, RajatおよびLajugie, JulienおよびJiang, YuおよびLi, HaochenおよびMargulies, Elliott、(2017)、Nirvana: Clinical Grade Variant Annotator、596-596、10.1145/3107411.3108204において説明される、Illumina Inc.によるNirvana(商標)アプリケーションなどのバリアントアノテーション/コールツールを利用することがあり、上記の文書の完全な主題の全体が、参照によって本明細書において明確に引用される。

【０２９３】

そのようなバリアントアノテーション/コールツールは、Nirvanaにおいて開示されるものなどの異なるアルゴリズム技法を適用することができる。

【０２９４】

a.区間アレイ(Interval Array)を用いてすべての重複する転写産物を特定する。機能的なアノテーションのために、バリアントと重複するすべての転写産物を特定することができ、区間木を使用することができる。しかしながら、区間のセットは静的であり得るので、区間木を区間アレイへとさらに最適化することが可能であった。区間木は、O(min(n,k lg n))時間においてすべての重複する転写産物を返し、ここでnは木の中の区間の数であ

り、kは重複する区間の数である。実際には、kは大半のバリアントに対してnと比較して本当に小さいので、区間木上での実効的なランタイムはO(k lg n)である。我々は、最初の重複する区間を見つけて、次いで残りの(k-1)にわたって数え上げるだけでよいように、ソートされたアレイにすべての区間が記憶されるような区間アレイを作成することによって、O(lg n+k)へと改善した。

【０２９５】

b.CNV/SV(Yu):コピー数変異および構造変異のためのアノテーションが提供され得る。小さいバリアントのアノテーションと同様に、SVと重複する転写産物、および以前に報告された構造変異も、オンラインデータベースにおいてアノテートされ得る。小さいバリアントと異なり、すべての重複する転写産物がアノテートされる必要はなく、それは、あまりにも多くの転写産物が大きなSVと重複するからである。代わりに、部分的な重複遺伝子に属するすべての重複する転写産物がアノテートされ得る。具体的には、これらの転写産物に対して、影響を受けるイントロン、エクソン、および構造変異により引き起こされる結果が報告され得る。すべての重複する転写産物の出力を許可するための選択肢が可能であるが、遺伝子名、転写産物との正規の(canonical)重複であるか部分的な重複であるかのフラグなどの、これらの転写産物の基本情報が報告され得る。各SV/CNVに対して、これらのバリアントが研究されているかどうか、および異なる集団におけるそれらの頻度を知ることも関心事である。したがって、1000 genomes、DGV、およびClinGenなどの外部データベースにおいて重複するSVを報告した。どのSVが重複しているかを決定するための恣意的なカットオフを使用するのを避けるために、代わりに、すべての重複する転写産物が使用されてよく、相互の重複率、すなわち、重複する長さをこれらの2つのSVの長さのうちの短い方で割ったものが計算されてよい。

【０２９６】

c.補足アノテーションを報告する。補足アノテーションには、小さいバリアントと構造バリアント(SV)という2つのタイプがある。SVは、区間としてモデル化されてよく、重複するSVを特定するために上で論じられた区間アレイを使用することができる。小さいバリアントは、点としてモデル化され、場所および(任意選択で)アレルによって照合される。したがって、それらは二分探索様のアルゴリズムを使用して検索される。補足アノテーションデータベースは非常に大きいことがあるので、補足アノテーションが存在するファイル位置に染色体場所をマッピングするために、はるかにより小さなインデックスが作成される。インデックスは、場所を使用して二分探索され得るオブジェクト(染色体場所とファイル位置からなる)のソートされたアレイである。インデックスサイズを小さく保つために、(ある最大のカウントまでの)複数の場所が、第1の場所の値と後続の場所に対する差分のみを記憶する1つのオブジェクトへと圧縮される。二分探索を使用するので、ランタイムはO(lg n)であり、nはデータベースの中の項目の数である。

【０２９７】

d.VEPキャッシュファイル

【０２９８】

e.転写産物データベース:転写産物キャッシュ(キャッシュ)および補足データベース(SAdb)ファイルは、転写産物および補足アノテーションなどのデータオブジェクトのシリアル化されたダンプである。キャッシュのためのデータソースとして、Ensembl VEPキャッシュを使用する。キャッシュを作成するために、すべての転写産物が区間アレイに挿入され、アレイの最終状態がキャッシュファイルに記憶される。したがって、アノテーションの間に、事前に計算された区間アレイをロードしてそれについて探索を実行するだけでよい。キャッシュはメモリへとロードされて探索は非常に高速(上で説明された)であるため、重複する転写産物を見つけることはNirvanaにおいては非常に高速である(総ランタイムの1%未満であると鑑定されている?)。

【０２９９】

f.補足データベース:SAdbのデータソースは補足材料のもとで列挙される。小さいバリアントに対するSAdbは、データベースの中の各オブジェクト(参照名および場所によって

特定される)がすべての関連する補足アノテーションを保持するように、すべてのデータソースのk-wayマージによって生み出される。データソースファイルを解析する間に遭遇する問題は、Nirvanaのホームページにおいて詳細に文書化されている。メモリ使用量を制限するために、SAインデックスのみがメモリにロードされる。このインデックスは、補足アノテーションのためのファイル位置の高速なルックアップを可能にする。しかしながら、データはディスクからフェッチされなければならないので、補足アノテーションを追加することは、Nirvanaの最大のボトルネックであると特定されている(総ランタイムの約30%であると鑑定されている)。

【０３００】

g.結果および配列オントロジー(Consequence and Sequence Ontology):Nirvanaの機能的アノテーション(提供されるとき)は、Sequence Ontology(SO)(http://www.sequenceontology.org/)ガイドラインに従う。時として、現在のSOにおける問題を特定して、アノテーションの状態を改善するためにSOチームと協力する機会があった。

【０３０１】

そのようなバリアントアノテーションツールは、前処理を含み得る。たとえば、Nirvanaは、ExAC、EVS、1000 Genomes project、dbSNP、ClinVar、Cosmic、DGV、およびClinGenのような、外部データソースからの多数のアノテーションを含んでいた。これらのデータベースを完全に利用するには、それらからの情報をサニタイジングしなければならない。我々は、様々なデータソースからの存在する様々な矛盾に対処するための異なる戦略を実施した。たとえば、同じ場所および代替のアレルに対する複数のdbSNPエントリがある場合、すべてのidをカンマで分けられたidのリストへと加える。同じアレルに対する異なるCAF値を伴う複数のエントリがある場合、第1のCAF値を使用する。矛盾するExACエントリとEVSエントリに対して、サンプルカウントの数を考慮し、よりサンプルカウントの高いエントリが使用される。1000 Genome Projectsでは、矛盾するアレルのアレル頻度を除去した。別の問題は不正確な情報である。我々は主に、1000 Genome Projectsからアレル頻度情報を抽出したが、GRCh38について、情報フィールドにおいて報告されているアレル頻度が利用可能ではない遺伝子型を伴うサンプルを除外していないことに気付き、これは、すべてのサンプルに対して利用可能ではないバリアントに対する頻度の低下につながる。我々のアノテーションの正確さを保証するために、個体レベルの遺伝子型のすべてを使用して真のアレル頻度を計算する。知られているように、同じバリアントは異なるアラインメントに基づく異なる表現を有し得る。すでに特定されたバリアントについての情報を正確に報告できることを確実にするために、異なるリソースからのバリアントを前処理してそれらを一貫した表現にしなければならない。すべての外部データソースに対して、基準アレルと代替アレルの両方における重複したヌクレオチドを除去するためにアレルをトリミングした。ClinVarについては、xmlファイルを直接解析し、すべてのバリアントに対する5'アラインメントを実行した。この手法はしばしばvcfファイルにおいて使用される。異なるデータベースは同じ情報のセットを含み得る。不必要な重複を避けるために、一部の重複した情報を除去した。たとえば、1000 genomesにおけるこれらのバリアントをより詳細な情報とともにすでに報告したので、1000 genome projectsをデータソースとして有するDGVの中のバリアントを除去した。

【０３０２】

少なくともいくつかの実装形態によれば、バリアントコールアプリケーションは、低頻度バリアント、生殖細胞系コーリングなどに対するコールを提供する。非限定的な例として、バリアントコールアプリケーションは、腫瘍のみのサンプルおよび/または腫瘍-正常ペアサンプルに対して実行され得る。バリアントコールアプリケーションは、一塩基変異(SNV)、多塩基変異(MNV)、インデルなどを探索し得る。バリアントコールアプリケーションは、バリアントを特定しながら、シーケンシングまたはサンプル調製エラーによる不一致をフィルタリングする。各バリアントに対して、バリアントコーラは、基準配列、バリアントの場所、および可能性のあるバリアント配列(たとえば、AからCへのSNV、またはAGからAへの欠失)を特定する。バリアントコールアプリケーションは、サンプル配列(また

はサンプルフラグメント)、基準配列/フラグメント、およびバリアントコールを、バリアントが存在することを示すものとして特定する。バリアントコールアプリケーションは、生フラグメントを特定し、生フラグメントの指定、潜在的なバリアントコールを検証する生フラグメントの数のカウント、支持バリアントが発生した生フラグメント内での場所、および他の関連する情報を特定し得る。生フラグメントの非限定的な例には、duplex stitchedフラグメント、simplex stitchedフラグメント、duplex un-stitchedフラグメント、およびsimplex un-stitchedフラグメントがある。

【０３０３】

バリアントコールアプリケーションは、.VCFファイルまたは.GVCFファイルなどの様々なフォーマットでコールを出力し得る。単なる例として、バリアントコールアプリケーションは、MiSeqReporterパイプラインに含まれ得る(たとえば、MiSeq(登録商標)シーケンサ装置で実装されるとき)。任意選択で、このアプリケーションは様々なワークフローを用いて実装され得る。分析は、所望の情報を得るために指定された方式でサンプルリードを分析する、単一のプロトコルまたはプロトコルの組合せを含み得る。

【０３０４】

次いで、1つまたは複数のプロセッサは、潜在的なバリアントコールに関連して妥当性確認操作を実行する。妥当性確認操作は、以下で説明されるように、品質スコア、および/または階層化された検定のヒエラルキーに基づき得る。妥当性確認操作が、潜在的なバリアントコールを確証または実証するとき、妥当性確認操作は(バリアントコールアプリケーションからの)バリアントコール情報をサンプル報告生成器に渡す。代わりに、妥当性確認操作が潜在的なバリアントコールを無効とするとき、または失格と判定するとき、妥当性確認操作は対応する指示(たとえば、否定的インジケータ、コールなしインジケータ、無効コールインジケータ)をサンプル報告生成器に渡す。妥当性確認操作はまた、バリアントコールが正しいことまたは無効なコール指定が正しいことの信頼性の程度に関する信頼性スコアを渡し得る。

【０３０５】

次に、1つまたは複数のプロセッサがサンプル報告を生成して記憶する。サンプル報告は、たとえば、サンプルに関する複数の遺伝子座に関する情報を含み得る。たとえば、遺伝子座の所定のセットの各遺伝子座に対して、サンプル報告は、遺伝子型コールを提供すること、遺伝子型コールを行えないことを示すこと、遺伝子型コールの確実さについての信頼性スコアを提供すること、または、1つまたは複数の遺伝子座に関するアッセイについての潜在的な問題を示すことのうちの少なくとも1つを行い得る。サンプル報告はまた、サンプルを提供した個体の性別を示し、かつ/またはサンプルが複数の源を含むことを示し得る。本明細書では、「サンプル報告」は、ある遺伝子座もしくは遺伝子座の所定のセットのデジタルデータ(たとえば、データファイル)および/または遺伝子座もしくは遺伝子座のセットの印刷された報告を含み得る。したがって、生成することまたは提供することは、データファイルを作成することおよび/もしくはサンプル報告を印刷すること、またはサンプル報告を表示することを含み得る。

【０３０６】

サンプル報告は、バリアントコールが決定されたが妥当性確認されなかったことを示すことがある。バリアントコールが無効であると決定されるとき、サンプル報告は、バリアントコールの妥当性を確認できないという決定の基礎に関する追加の情報を示し得る。たとえば、報告の中の追加の情報は、生フラグメントの記述と、生フラグメントがバリアントコールを支持または否定する程度(たとえば、カウント)とを含み得る。加えて、または代わりに、報告の中の追加の情報は、本明細書で説明される実装形態に従って得られる品質スコアを含み得る。

【０３０７】

［バリアントコールアプリケーション］

本明細書で開示される実装形態は、潜在的なバリアントコールを特定するためにシーケンシングデータを分析することを含む。バリアントコールは、以前に実行されたシーケン

シング操作について記憶されたデータに対して実行され得る。加えて、または代わりに、バリアントコーリングは、シーケンシング操作が実行されている間にリアルタイムで実行され得る。サンプルリードの各々が、対応する遺伝子座を割り当てられる。サンプルリードは、サンプルリードのヌクレオチドの配列、または言い換えると、サンプルリード内のヌクレオチドの順序(たとえば、A、C、G、T)に基づいて、対応する遺伝子座に割り当てられ得る。この分析に基づいて、サンプルリードは、特定の遺伝子座の潜在的なバリアント/アレルを含むものとして指定され得る。サンプルリードは、遺伝子座の潜在的なバリアント/アレルを含むものとして指定された他のサンプルリードとともに収集(または集約または貯蔵)され得る。割当て操作はコーリング操作とも呼ばれることがあり、コーリング操作において、サンプルリードは特定の遺伝子場所/座と関連付けられる可能性があるものとして特定される。サンプルリードは、サンプルリードを他のサンプルリードから区別するヌクレオチドの1つまたは複数の識別配列(たとえば、プライマー配列)を位置特定するために分析され得る。より具体的には、識別配列は、特定の遺伝子座と関連付けられるものとしてサンプルリードを他のサンプルリードから特定し得る。

【０３０８】

割当て操作は、識別配列の一連のn個のヌクレオチドが選択配列のうちの1つまたは複数と実質的に一致するかどうかを決定するために、識別配列の一連のn個のヌクレオチドを分析することを含み得る。特定の実装形態では、割当て操作は、サンプル配列の最初のn個のヌクレオチドが選択配列のうちの1つまたは複数と実質的に一致するかどうかを決定するために、サンプル配列の最初のn個のヌクレオチドを分析することを含み得る。数nは様々な値を有することがあり、この値はプロトコルへとプログラムされることがあり、またはユーザにより入力されることがある。たとえば、数nは、データベース内の最短の選択配列のヌクレオチドの数として定義され得る。数nは所定の数であり得る。所定の数は、たとえば、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、または30個のヌクレオチドであり得る。しかしながら、他の実装形態では、より少数または多数のヌクレオチドが使用され得る。数nはまた、システムのユーザなどの個人によって選択されてもよい。数nは1つまたは複数の条件に基づき得る。たとえば、数nは、データベース内の最短のプライマー配列のヌクレオチドの数または指定された数のうちの小さい方の数として定義され得る。いくつかの実装形態では、15個未満のヌクレオチドであるあらゆるプライマー配列が例外として指定され得るように、15などのnの最小値が使用され得る。

【０３０９】

いくつかの場合、識別配列の一連のn個のヌクレオチドは、選択配列のヌクレオチドと正確に一致しないことがある。それでも、識別配列は、選択配列とほぼ同一である場合、選択配列と実質的に一致し得る。たとえば、識別配列の一連のn個のヌクレオチド(たとえば、最初のn個のヌクレオチド)が、不一致が指定された数(たとえば、3)を超えずに、かつ/またはシフトが指定された数(たとえば、2)を超えずに、選択配列と一致する場合、サンプルリードが遺伝子座に対してコールされ得る。各不一致またはシフトが、サンプルリードとプライマー配列との間の差としてカウントされ得るように、規則が確立され得る。差の数が指定された数未満である場合、サンプルリードは、対応する遺伝子座(すなわち、対応する遺伝子座に割り当てられる)に対してコールされ得る。いくつかの実装形態では、サンプルリードの識別配列と遺伝子座に関連付けられる選択配列との間の差の数に基づく、マッチングスコアが決定され得る。マッチングスコアが指定されたマッチング閾値に合格する場合、選択配列に対応する遺伝子座は、サンプルリードに対する潜在的な座として指定され得る。いくつかの実装形態では、サンプルリードが遺伝子座に対してコールされるかどうかを決定するために、後続の分析が実行され得る。

【０３１０】

サンプルリードがデータベースの中の選択配列のうちの1つと実質的に一致する(すなわち、厳密に一致する、または上で説明されたようにほぼ一致する)場合、サンプルリードは、選択配列と相関する遺伝子座に割り当てられ、または指定される。これは、座コーリ

ングまたは予備的座コーリングと呼ばれることがあり、選択配列に相関する遺伝子座に対してサンプルリードがコールされる。しかしながら、上で論じられたように、サンプルリードは2つ以上の遺伝子座に対してコールされ得る。そのような実装形態では、潜在的な遺伝子座のうちの1つだけに対するサンプルリードをコールするために、または割り当てるために、さらなる分析が実行され得る。いくつかの実装形態では、基準配列のデータベースと比較されるサンプルリードは、ペアエンドシーケンシングからの最初のリードである。ペアエンドシーケンシングを実行するとき、サンプルリードに相関する第2のリード(生フラグメントを表す)が得られる。割当ての後で、割り当てられたリードについて実行される後続の分析は、割り当てられたリードに対してコールされた遺伝子座のタイプに基づき得る。

【０３１１】

次に、潜在的なバリアントコールを特定するために、サンプルリードが分析される。とりわけ、分析の結果は、潜在的なバリアントコール、サンプルバリアント頻度、基準配列、および対象のゲノム配列内でのバリアントが発生した場所を特定する。たとえば、遺伝子座がSNPを含むことが知られている場合、遺伝子座に対してコールされた割り当てられたリードは、割り当てられたリードのSNPを特定するための分析を経ることがある。遺伝子座が多型の反復的なDNA要素を含むことが知られている場合、サンプルリード内の多型の反復的なDNA要素を特定するために、または特徴付けるために、割り当てられるリードが分析され得る。いくつかの実装形態では、割り当てられるリードがSTR座およびSNP座と実質的に一致する場合、警告またはフラグがサンプルリードに割り当てられ得る。サンプルリードは、STR座とSNP座の両方として指定され得る。この分析は、割り当てられたリードの配列および/または長さを決定するために、割当てプロトコルに従って割り当てられたリードをアラインメントすることを含み得る。アラインメントプロトコルは、全体が参照によって本明細書において引用される、2013年3月15日に出願された国際特許出願第PCT/US2013/030867号(公開番号第WO 2014/142831号)において説明される方法を含み得る。

【０３１２】

次いで、1つまたは複数のプロセッサは、支持バリアントが生フラグメント内の対応する場所に存在するかどうかを決定するために、生フラグメントを分析する。様々なタイプの生フラグメントが特定され得る。たとえば、バリアントコーラは、元のバリアントコールを妥当性確認するバリアントを示すあるタイプの生フラグメントを特定し得る。たとえば、そのタイプの生フラグメントは、duplex stitchedフラグメント、simplex stitchedフラグメント、duplex un-stitchedフラグメント、またはsimplex un-stitchedフラグメントを表し得る。任意選択で、前述の例の代わりに、またはそれに加えて、他の生フラグメントが特定され得る。各タイプの生フラグメントを特定することに関連して、バリアントコーラはまた、支持バリアントが発生した生フラグメント内での場所、ならびに、支持バリアントを示した生フラグメントの数のカウントを特定する。たとえば、バリアントコーラは、生フラグメントの10個のリードが、特定の場所Xにおいて支持バリアントを有するduplex stitchedフラグメントを表すことが特定されたことを示すものを、出力し得る。バリアントコーラはまた、生フラグメントの5個のリードが、特定の場所Yにおいて支持バリアントを有するsimplex un-stitchedフラグメントを表すことが特定されたことを示すものを、出力し得る。バリアントコーラはまた、基準配列に対応し、したがって対象のゲノム配列における潜在的なバリアントコールを妥当性確認する証拠を提供する支持バリアントを含まなかった、生フラグメントの数を出力し得る。

【０３１３】

次に、支持バリアント、ならびに支持バリアントが発生した場所を含む、生フラグメントのカウントが維持される。加えて、または代わりに、(サンプルリードまたはサンプルフラグメントの中の潜在的なバリアントコールの場所に対する相対的な)対象の場所において支持バリアントを含まなかった生フラグメントのカウントが維持され得る。加えて、または代わりに、基準配列に対応し潜在的なバリアントコールを確証または確認しない、生フラグメントのカウントが維持され得る。潜在的なバリアントコールを支持する生フラ

グメントのカウントおよびタイプ、生フラグメントの中の支持バリアントの場所、潜在的なバリアントコールを支持しない生フラグメントのカウントなどを含む、決定された情報が、バリアントコール妥当性確認アプリケーションに出力される。

【０３１４】

潜在的なバリアントコールが特定されるとき、プロセスは、潜在的なバリアントコール、バリアント配列、バリアント場所、およびそれらと関連付けられる基準配列を示すものを出力する。エラーはコールプロセスに誤ったバリアントを特定させ得るので、バリアントコールは「潜在的な」バリアントを表すように指定される。本明細書の実装形態によれば、誤ったバリアントまたは偽陽性を減らして除去するために、潜在的なバリアントコールが分析される。加えて、または代わりに、プロセスは、サンプルリードと関連付けられる1つまたは複数の生フラグメントを分析し、生フラグメントと関連付けられる対応するバリアントコールを出力する。

【０３１５】

［良性訓練セットの生成］

数百万個のヒトゲノムおよびエクソンがシーケンシングされているが、それらの臨床上の応用は、疾患を引き起こす変異を良性の遺伝的変異から区別することの難しさにより限られたままである。ここで我々は、他の霊長類の種における一般的なミスセンスバリアントが、ヒトにおいて大部分が臨床的に良性であることを実証し、病原性の変異が除去のプロセスによって系統的に特定されることを可能にする。6種のヒト以外の霊長類の種の集団シーケンシングからの数十万個の一般的なバリアントを使用して、88%の正確さで稀な疾患の患者における病原性の変異を特定し、ゲノムワイド有意性(genome-wide significance)で知的障害における14個の新たな遺伝子候補の発見を可能にする、深層ニューラルネットワークを訓練した。追加の霊長類の種からの一般的な変異の目録を作ることで、数百万個の有意性が不確かなバリアントに対する解釈が改善し、ヒトゲノムシーケンシングの臨床上の利用がさらに進む。

【０３１６】

診断シーケンシングの臨床上の使用可能性は、ヒトの集団における稀な遺伝子バリアントを解釈しそれらの疾患リスクに対する影響を推測することが難しいことにより、限られている。臨床的に有意な遺伝子バリアントは、それらの健康に対する有害な影響により、集団において極めて稀である傾向があり、大半については、ヒトの健康に対する影響が決定されていない。臨床的な有意性が不確かであるこれらのバリアントが多数あること、およびそれらが稀であることは、個人化された医療および集団全体の健康スクリーニングに対するシーケンシングの採用に対する手強い障壁となっている。

【０３１７】

大半の浸透性のメンデル性の疾患は集団において非常に有病率が低いので、集団における高頻度でのバリアントの観察は、良性の結果を支持する強い証拠である。多様なヒトの集団にわたって一般的な変異を評価することは、良性のバリアントの目録を作るための有効な戦略であるが、現生人類における一般的な変異の総数は、祖先の多様性の大部分が失われた我々の種の最近の歴史におけるボトルネック事象により、限られている。現生人類の集団の研究は、過去15000～65000年以内の10000人未満の個人という有効個体数(N_e)からの顕著な膨張を示しており、一般的な多型のプールが小さいことは、このサイズの集団における変異の容量が限られていることに由来する。基準ゲノムの中の7000万個の潜在的なタンパク質を変化させるミスセンス置換のうち、全体で0.1%を超える集団アレル頻度を持つものは、概ね1000個のうちの1個しか存在しない。

【０３１８】

現生人類の集団以外では、チンパンジーが次に近い現存する種を構成し、99.4%のアミノ酸配列相同性を共有する。ヒトとチンパンジーにおけるタンパク質コーディング配列の近い相同性は、チンパンジーのタンパク質コーディングバリアントに対して作用する純化選択が、同一状態であるヒトの変異の健康に対する結果もモデル化し得ることを示唆する。

【０３１９】
中立的な多型がヒトの祖先の系統(約4N_e世代)において持続する平均時間は、種の分岐時間(約600万年前)の一部であるので、自然に発生するチンパンジーの変異は、平衡選択により維持されるハプロタイプの稀な事例を除き、偶然を除いて大部分が重複しない変異空間に及ぶ。同一状態である多型が2つの種において同様に健康に影響する場合、チンパンジーの集団における高いアレル頻度でのバリアントの存在は、ヒトにおける良性の結果を示すはずであり、その良性の結果が純化選択によって確立されている既知のバリアントの目録を拡大する。
【０３２０】
［結果-他の霊長類における一般的なバリアントはヒトにおいて大部分が良性である］
Exome Aggregation Consortium(ExAC)およびGenome Aggregation Database(gnomAD)において収集された123136人のヒトを含む、集約されたエクソンデータが最近利用可能になったことで、アレル頻度スペクトラムにわたるミスセンス変異と同義変異に対する自然選択の影響を測ることが可能になった。コホートにおいて1回しか観察されない稀なシングルトンバリアントは、変異率に対するトリヌクレオチドコンテクストの影響を調整した後の、de novo変異によって予測される、予想される2.2/1のミスセンス/同義比とよく一致する(図49A、図51、ならびに図52A、図52B、図52C、および図52D)が、より高いアレル頻度では、観察されるミスセンスバリアントの数は、自然選択による有害な変異の一掃により減少する。アレル頻度の増大に伴うミスセンス/同義比の段階的な低下は、集団頻度が<0.1%であるミスセンスバリアントのかなりの部分が、健康な個人において観察されるにもかかわらず軽度に有害な結果を有することと一致する。これらの発見は、0.1%～約1%より高いアレル頻度を伴うバリアントを、平衡選択および創始者効果により引き起こされるよく記録されている少数の例外を除いて、浸透性の遺伝性疾患に対しては良性である可能性が高いものとして除去するという、診療室において広く行われている経験的な実践を支持するものである。
【０３２１】
我々は、24体の親類ではない個体のコホートにおいて2回以上サンプリングされた、一般的なチンパンジーバリアントを特定した。これらのバリアントの99.8%が一般のチンパンジー集団において一般的である(アレル頻度(AF)>0.1%)ことが推定され、これは、これらのバリアントが純化選択のふるいにすでにかけられていることを示す。我々は、複数の配列アラインメントにおける1対1のマッピングを欠いているバリアントとともに、延長された主要組織適合遺伝子複合体領域を平衡選択の既知の領域として除いて、対応する同一状態のヒトバリアントに対するヒトアレル頻度スペクトラムを調査した(図49B)。一般的なチンパンジーバリアントと同一状態であるヒトバリアントについて、ミスセンス/同義比はヒトアレル頻度スペクトラムにわたって概ね一定であり(カイ二乗($\chi 2$)検定によりP>0.5)、これは、ヒトの集団における、一般的なチンパンジーバリアントに対する負の選択がないことと、2つの種におけるミスセンスバリアントに対する選択係数が合致していることと一致する。一般的なチンパンジーバリアントと同一状態であるヒトバリアントにおいて観察される低いミスセンス/同義比は、チンパンジーのより大きな有効個体数(N_e約73000)と一致し、これは軽度に有害な変異のより効率的な除去を可能にする。
【０３２２】
対照的に、シングルトンチンパンジーバリアント(コホートにおいて1回しかサンプリングされない)について、一般的なアレル頻度においてミスセンス/同義比の大幅な低下が観察され($P<5.8\times10^{-6}$、図49C)、これは、シングルトンチンパンジーミスセンスバリアントの24%が、0.1%より高いアレル頻度ではヒトの集団における純化選択によってフィルタリングされるであろうことを示している。この枯渇は、チンパンジーシングルトンバリアントの大部分が、その健康に対する有害な影響によりいずれの種においても一般的なアレル頻度に達することが妨げられた、稀な有害変異であることを示している。我々は、シングルトンバリアントの69%だけが、一般のチンパンジー集団において一般的である(AF>0.1%)と推定する。

【０３２３】
次に、6種のヒト以外の霊長類の種のうちの少なくとも1種において観察される変異と同一状態であるヒトバリアントを特定した。6種の各々における変異は、大型類人猿ゲノムプロジェクト(チンパンジー、ボノボ、ゴリラ、およびオランウータン)から確認され、または、霊長類ゲノムプロジェクト(アカゲザル、マーモセット)から一塩基多型データベース(dbSNP)に提出され、シーケンシングされた個体の数が限られていること、および各種について観察されたミスセンス:同義比(補足テーブル1)が低いことに基づいて、大部分が一般的なバリアントを表す。チンパンジーと同様に、6種のヒト以外の霊長類の種からのバリアントのミスセンス/同義比は、少ない割合(チンパンジーにおいて0.1%未満のアレル頻度、および他の種ではシーケンシングされた個体がより少ないことによってより低いアレル頻度のもとで、約16%)の稀なバリアントが含まれることにより予想される一般的なアレル頻度におけるミスセンス変異の軽度の枯渇(図49D、図53、図54、および図55、補足データファイル1)を除き、ヒトアレル頻度スペクトラムにわたって概ね等しいことを発見した。これらの結果は、同一状態のミスセンスバリアントに対する選択係数が、ヒトの祖先の系統から約3500万年前に分岐したと推定される少なくとも新世界ザルまでの霊長類の系統内で合致していることを示唆する。

【０３２４】
観察された霊長類のバリアントと同一状態であるヒトミスセンスバリアントは、ClinVarデータベースにおける良性の結果に対して強くエンリッチされることを発見した。有意性が不確かであるバリアントおよび矛盾するアノテーションを伴うバリアントを除外した後で、少なくとも1種のヒト以外の霊長類の種において存在するClinVarバリアントは、平均で90%の事例で良性または良性である可能性が高いものとしてアノテートされ、それと比較して、ClinVarミスセンスバリアント全般では35%である($P<10^{-40}$、図49E)。霊長類バリアントに対するClinVarアノテーションの病原性は、精選バイアスを減らすために1%より高いアレル頻度を伴うヒトバリアントを除く健康なヒトの同様のサイズのコホートをサンプリングすることから観察されるもの(約95%が良性または良性である可能性が高い結果、P=0.07)より、わずかに高い。

【０３２５】
ヒトの遺伝学の分野は、ヒト変異の臨床上の影響を推論するためにモデル生物に長い間依存してきたが、大半の遺伝的に扱いやすい動物モデルまでの進化的距離が長いことで、モデル生物についての発見がヒトに対してどの程度一般化可能であるかについての懸念が生まれている。我々は、4種の追加の哺乳類の種(ネズミ、ブタ、ヤギ、ウシ)と2種のより遠縁の脊椎動物(ニワトリ、ゼブラフィッシュ)からの概ね一般的な変異を含めるように、霊長類の系統を超えて分析を拡張した。我々は、dbSNPにおいてゲノムワイドの変異の確認が十分にとれている種を選択し、これらが概ね一般的なバリアントであることを、ミスセンス/同義比が2.2/1よりはるかに小さいことに基づいて確認した。我々の霊長類の分析とは対照的に、より遠縁の種における変異と同一状態であるヒトミスセンス変異は、一般的なアレル頻度では顕著に枯渇しており(図50A)、この枯渇の程度はより長い進化的距離において増大する(図50Bおよび補足テーブル2および3)。

【０３２６】
ヒトにおいて有害であるが、より遠縁の種では高いアレル頻度で耐えているミスセンス変異は、同一状態のミスセンス変異に対する選択の係数が、ヒトとより遠縁の種との間でかなり離れていることを示す。それでも、より遠縁の哺乳類におけるミスセンスバリアントの存在は良性の結果の確率を高め、それは、一般的なアレル頻度において自然選択により枯渇するミスセンスバリアントの割合が、ヒトミスセンスバリアント全般について観察される約50%の枯渇率より低いからである(図49A)。これらの結果と一致して、ネズミ、ブタ、ヤギ、およびウシにおいて観察されているClinVarミスセンスバリアントは、良性の結果または良性である可能性が高い結果をアノテートされる確率が73%であり、それと比較して、霊長類の変異に対しては90%であり($P<2\times10^{-8}$、図50C)、ClinVarデータベース全体に対しては35%であることを発見した。

【０３２７】
家畜化によるアーティファクトではなく進化的距離が選択係数の相違の主な原因であることを確認するために、我々は、広範囲の進化的距離にわたって、種内多型の代わりに近縁の種のペア間での固定された置換を使用して、分析を繰り返した(図50D、補足テーブル4、および補足データファイル2)。我々は、種間の固定された置換と同一状態であるヒトミスセンスバリアントの枯渇率が、進化的な枝長とともに増大し、家畜化を受けた種と比較して野生種に対する識別可能な差がないことを発見した。これは、同一状態の固定されたミスセンス置換の数が分岐した系統において偶然により予想されるものよりも低かったことを発見した、ハエおよび酵母菌における以前の成果と一致している。

【０３２８】
［バリアントの病原性分類のための深層学習ネットワーク］
開示される技術は、バリアントの病原性分類のための深層学習ネットワークを提供する。臨床上の応用に対するバリアント分類の重要性は、教師あり機械学習を問題の対処のために使用する多くの試みを引き起こしてきたが、これらの努力は、訓練のために確信をもってラベリングされた良性のバリアントおよび病原性のバリアントを含む適切なサイズの真実データセット(truth dataset)がないことにより、妨げられている。

【０３２９】
専門家により精選されたバリアントの既存のデータベースはゲノム全体を代表しておらず、ClinVarデータベースの中のバリアントの約50%がわずか200個の遺伝子(ヒトのタンパク質コーディング遺伝子の約1%)に由来する。その上、系統的な研究により、多くの専門家のアノテーションには証拠が疑わしいものがあることが特定されており、単一の患者においてのみ観察され得る稀なバリアントを解釈することの難しさを示している。専門家の解釈はますます厳密になっているが、分類のガイドラインは、大部分が合意された習慣に沿って策定されており、既存の傾向を強めるリスクがある。人による解釈のバイアスを減らすために、最近に分類器は、一般的なヒト多型または固定されたヒト-チンパンジーの置換に対して訓練されているが、これらの分類器も、人により精選されたデータベース上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している。これらの様々な方法の性能の客観的なベンチマーキングは、独立でバイアスのない真実データセットがなければ、達成が難しい。

【０３３０】
6種のヒト以外の霊長類(チンパンジー、ボノボ、ゴリラ、オランウータン、アカゲザル、およびマーモセット)からの変異は、一般的なヒト変異と重複しない300000個を超える固有のミスセンスバリアントに寄与し、大部分が純化選択のふるいにかけられた良性の結果の一般的なバリアントを表し、機械学習の手法に利用可能な訓練データセットを大きく拡大する。平均すると、各霊長類の種は、ClinVarデータベースの全体より多くのバリアント(2017年11月現在で、有意性が不確かなバリアントおよび矛盾するアノテーションを伴うバリアントを除いた後で、約42000個のミスセンスバリアント)に寄与する。加えて、この内容は人の解釈によるバイアスがない。

【０３３１】
一般的なヒトバリアント(AF>0.1%)および霊長類の変異(補足テーブル5(図58))を備えるデータセットを使用して、我々は新しい深層残差ネットワークPrimateAIを訓練した。PrimateAIは、対象のバリアントの側にあるアミノ酸配列および他の種におけるオーソロガス配列アラインメントを入力として取り込む(図2および図3)。人により加工された特徴量を利用する既存の分類器と異なり、我々の深層学習ネットワークは、元の配列から直接特徴量を抽出するように学習する。タンパク質構造についての情報を組み込むために、二次構造および溶媒接触性を配列だけから予測するように別々のネットワークを訓練し、次いでこれらを完全なモデルにおけるサブネットワークとして含めた(図5および図6)。結晶化に成功している少数のヒトタンパク質を仮定すれば、元の配列から構造を推論することには、不完全なタンパク質構造および機能ドメインのアノテーションによるバイアスが避けられるという利点がある。ネットワークの全体の深さは、含まれるタンパク質構造とともに

、およそ400000個の訓練可能なパラメータを備える36層の畳み込みであった。

【０３３２】

良性のラベルを伴うバリアントのみを使用して分類器を訓練するために、我々は、所与の変異が集団において一般的なバリアントとして観察される可能性が高いかどうかということとして、予測問題を形作った。いくつかの要因が高いアレル頻度でのバリアントの観察の確率に影響し、我々はそれらのバリアントの有害性だけに関心がある。他の要因には、変異率、シーケンシングカバレッジなどの技術的なアーティファクト、および遺伝子変換などの中立的な遺伝的浮動に影響する要因がある。

【０３３３】

我々は、これらの区別できない要因の各々を考慮して、良性の訓練セットの中の各バリアントをExACデータベースからの123136個のエクソンに存在しなかったミスセンス変異と照合し、良性のバリアントと照合された対照群とを区別するように深層学習ネットワークを訓練した(図24)。ラベリングされていないバリアントの数は、ラベリングされた良性の訓練データセットのサイズを大きく超えるので、良性の訓練データセットと照合されたラベリングされていないバリアントの異なるセットを各々使用して、8個のネットワークを並列に訓練し、コンセンサス予測を得た。

【０３３４】

元のアミノ酸配列のみを入力として使用して、深層学習ネットワークは、てんかん、自閉症、および知的障害における主要な疾患遺伝子である電位依存性ナトリウムチャネルSCN2Aについて示されるように(図20)、有用なタンパク質機能ドメインにおける残基に高い病原性スコアを正確に割り当てている。SCN2Aの構造は、各々が6つの膜貫通ヘリックス(S1～S6)を含む4つの相同なリピートを備える。膜の脱分極により、正に荷電したS4膜貫通ヘリックスが膜の細胞外の側に向かって動き、S4-S5リンカーを介してS5/S6孔形成ドメインを開口させる。てんかん性脳症の早期の兆候と臨床的に関連付けられる、S4、S4-S5リンカー、およびS5ドメインにおける変異は、遺伝子において最高の病原性スコアを有するものとしてネットワークにより予測され、健康な集団におけるバリアントに対して枯渇している(補足テーブル6)。我々はまた、ネットワークが、ドメイン内の重要なアミノ酸の場所を認識し、転写因子のDNA接触残基および酵素の触媒残基などの、これらの場所における変異に最高の病原性スコアを割り当てていることも発見した(図25A、図25B、図25C、および図26)。

【０３３５】

深層学習ネットワークがタンパク質の構造および機能についての洞察を元の配列からどれだけ導いているかをより理解するために、ネットワークの最初の3層からの訓練可能なパラメータを視覚化した。これらの層内で、Granthamスコアなどのアミノ酸距離の既存の測定結果に近い、異なるアミノ酸の重みと重みの間の相関を、ネットワークが学習することが観察された(図27)。これらの初期の層の出力はより後の層の入力になり、深層学習ネットワークがデータの次第により高次の表現を構築することを可能にする。

【０３３６】

訓練を保留された10000個の一般的な霊長類バリアントを使用して、既存の分類アルゴリズムを用いたネットワークの性能を比較した。すべての新たに生じたヒトミスセンスバリアントの約50%が一般的なアレル頻度で純化選択によってフィルタリングされるので(図49A)、変異率およびシーケンシングカバレッジによって10000個の一般的な霊長類バリアントと照合された10000個のランダムに選択されたバリアントのセットに対する各分類器の50パーセンタイルスコアを求め、その閾値における各分類器の正確さを評価した(図21D、図28A、および補足データファイル4)。10000個の保留された一般的な霊長類バリアントに良性の結果を割り当てることについて、我々の深層学習ネットワーク(91%の正確さ)は他の分類器の性能(次に良いモデルで80%の正確さ)を上回った。

【０３３７】

ヒト変異データのみを用いて訓練されたネットワークの正確さ(図21D)と比較した場合、既存の方法を超える改善の概ね半分は、深層学習ネットワークを使用することに由来し

、半分は訓練データセットを霊長類の変異で補強することに由来する。ある臨床シナリオにおける有意性が不確かなバリアントの分類を検定するために、神経発達障害の患者vs健康な対照群において発生するde novo変異を区別する、深層学習ネットワークの能力を評価した。有病率により、神経発達障害は稀な遺伝子疾患の最大のカテゴリのうちの1つを構成しており、最近のトリオシーケンシング研究は、de novoミスセンス変異およびタンパク質切断変異の中心的な役割を示唆している。

【０３３８】

各々確信をもってコールされている、Deciphering Developmental Disorders(DDD)コホートからの4293人の影響を受けている個人におけるde novoミスセンスバリアントvs Simon's Simplex Collection(SSC)コホートにおける2517人の影響を受けていない兄弟からのde novoミスセンスバリアントを分類し、ウィルコクソンの順位和検定を用いて2つの分布の間での予測スコアの差を評価した(図21Eおよび図29Aおよび図29B)。深層学習ネットワークは、このタスクについて他の分類器を明確に上回った($P<10^{-28}$、図21Fおよび図28B)。その上、保留された霊長類バリアントデータセットと、DDD症例群vs対照群データセットとに対する、様々な分類器の性能が相関付けられ(スピアマンρ=0.57、P<0.01)、全く異なる源および方法を使用しているにもかかわらず、病原性の評価について2つのデータセットの間で良好な一致を示した(図30A)。

【０３３９】

次に、同じ遺伝子内で良性変異vs病原性変異を分類することについての、深層学習ネットワークの正確さを推定することを試みる。DDD集団の大部分が、影響を受けている第一度近親者のいない、影響を受けている子供のインデックスケースを備えると仮定すると、分類器がde novo優性遺伝モードを持つ遺伝子の病原性を過剰評価することによって正確さを釣り上げていないことを示すのが重要である。我々は、タンパク質切断変異(P<0.05)だけから計算された、DDD研究において疾患との関連について名目上有意であった605個の遺伝子に分析を制約した。これらの遺伝子内で、de novoミスセンス変異は、予想と比較して3/1エンリッチされており(図22A)、約67%が病原性であることを示している。

【０３４０】

深層学習ネットワークは、遺伝子の同じセット内で病原性のde novoバリアントと良性のde novoバリアントを区別することが可能であり($P<10^{-15}$、図22B)、他の方法の性能を大きく上回った(図22Cおよび図28C)。0.803以上というバイナリカットオフでは(図22Dおよび図30B)、症例群におけるde novoミスセンス変異の65%が病原性であるものとして深層学習ネットワークにより分類され、それと比較して、対照群においてはde novoミスセンス変異の14%が病原性であり、これは88%という分類の正確さに対応する(図22Eおよび図30C)。神経発達障害における頻繁な不完全な浸透性および変化する表現性を考慮すると、この数字は、対照群において部分的に浸透している病原性バリアントが含まれていることにより、我々の分類器の正確さをおそらく過小評価している。

【０３４１】

［新しい遺伝子候補の発見］

病原性ミスセンス変異を階層化するために0.803以上の閾値を適用することは、DDD患者におけるde novoミスセンス変異のエンリッチメントを、1.5-foldからタンパク質切断変異(2.5-fold)に近い2.2-foldへと増大させ、一方で、予想を超えてエンリッチされるバリアントの総数の3分の1未満を捨てる。このことは、統計能力をかなり高め、元のDDD研究ではゲノムワイド有意性閾値にこれまで達していなかった知的障害における14個の追加の遺伝子候補の発見を可能にしている(テーブル1)。

【０３４２】

［専門家による精選との比較］

ClinVarデータベースからの最近の専門家により精選されたバリアントに対する様々な分類器の性能を調査したが、ClinVarデータセットに対する分類器の性能は、保留された霊長類バリアントデータセットとも、DDD症例群vs対照群データセットとも強く相関していなかったことを発見した(それぞれP=0.12およびP=0.34)(図31Aおよび図31B)。我々は、

既存の分類器には専門家の精選によるバイアスがあるという仮説を立てており、人の経験則は正しい方向にある傾向にあるものの最適ではないことがある。1つの例は、ClinVarにおける病原性バリアントと良性バリアントとの間のGranthamスコアの平均の差であり、これは、605個の疾患関連遺伝子内での、DDD症例群vs対照群におけるde novoバリアントの差の2倍である(テーブル2)。それと比べて、専門家による精選は、タンパク質構造を、特に、他の分子と相互作用することが可能になり得る表面に曝露されている残基の重要性を、十分に活用していないように見える。我々は、ClinVar病原性変異とDDD de novo変異の両方が、予測される溶媒に曝露される残基と関連付けられるが、良性のClinVarバリアントと病原性のClinVarバリアントとの間の溶媒接触性の差はDDD症例群vs対照群について見られる差の半分にすぎないことを観察した。これらの発見は、Granthamスコアおよび保存率などの、専門家にとって解釈がより簡単な要因を優先する確認バイアスを示唆するものである。人により精選されたデータベース上で訓練された機械学習分類器は、これらの傾向を強化することが予想される。

【０３４３】

我々の結果は、系統的な霊長類集団のシーケンシングが、臨床上のゲノム解釈を現在制約している、数百万個の有意性が不確かなヒトバリアントを分類するための有効な戦略であることを示唆している。保留された一般的な霊長類バリアントと臨床上のバリアントの両方に対する我々の深層学習ネットワークの正確さは、ネットワークを訓練するために使用される良性バリアントの数とともに高まる(図23A)。その上、6種のヒト以外の霊長類の種の各々からのバリアントについての訓練は、ネットワークの性能の向上に独立に寄与し、一方で、より遠縁の哺乳類からのバリアントについての訓練はネットワークの性能に負の影響を与える(図23Bおよび図23C)。これらの結果は、一般的な霊長類バリアントが、浸透性のメンデル性疾患に関してヒトにおいて大部分が良性である一方で、より遠縁の種における変異については同じことが言えないという主張を、支持している。

【０３４４】

本研究において調査されるヒト以外の霊長類ゲノムの数は、シーケンシングされてきたヒトゲノムおよびエクソンの数と比較して少ないが、これらの追加の霊長類は、一般的な良性変異についての不相応な量の情報に寄与していることに留意することが重要である。ExACを用いたシミュレーションは、一般的なヒトバリアント(>0.1%アレル頻度)の発見はわずか数百の個体の後ですぐに停滞するが(図56)、数百万人までのさらなる健康な集団のシーケンシングが主に追加の稀なバリアントに寄与することを示している。アレル頻度に基づいて大部分が臨床的に良性であることが知られている一般的なバリアントと異なり、健康な集団における稀なバリアントは、劣性の遺伝的疾患または浸透性が不完全である優性の遺伝的疾患を引き起こし得る。各霊長類の種は一般的なバリアントの異なるプールを持つので、各種の数十の個体をシーケンシングすることは、霊長類の系統における良性ミスセンス変異の目録を系統的に作るのに有効な戦略である。実際に、本研究で調査された6種のヒト以外の霊長類の種からの134の個体が、ExAC研究からの123136人のヒトの4倍近く多くの一般的なミスセンス変異に寄与している(補足テーブル5(図58))。数百の個体に関する霊長類集団のシーケンシング研究は、野生の保護区域および動物園に住む比較的少数の親類ではない個体でも現実的であり得るので、野生の集団に対する外乱が最小限になり、これはヒト以外の霊長類の保全および倫理的な取り扱いの観点から重要である。

【０３４５】

現生人類の集団は、大半のヒト以外の霊長類の種よりはるかに遺伝的な多様性が低く、チンパンジー、ゴリラ、およびテナガザルと比べて、個体ごとの一塩基バリアントの数が概ね半分であり、オランウータンと比べて個体ごとのバリアントが3分の1である。ヒト以外の霊長類の種の大半の遺伝的多様性のレベルは知られていないが、多数の現存するヒト以外の霊長類の種により、潜在的な良性のヒトミスセンスの場所の大半が少なくとも1つの霊長類の種における一般的なバリアントによってカバーされる可能性が高いと推定することが可能になり、病原性バリアントが除去のプロセスによって系統的に特定されることが可能になる(図23D)。シーケンシングされるこれらの種のサブセットのみでも、訓練デ

ータサイズを大きくすることで、機械学習を用いたミスセンスの結果のより正確な予測が可能になる。最終的に、我々の発見はミスセンス変異に注目しているが、この戦略は、バリアントが同一状態であるかどうかを明確に決定するための十分なアラインメントがヒトゲノムと霊長類ゲノムとの間にある保存された制御領域においては特に、非コーディング変異の結果を推論することにも適用可能であり得る。

【０３４６】

504種の既知のヒト以外の霊長類の種のうち、約60%が密猟および大規模な生息地の喪失により絶滅に瀕している。これらの種における個体数の減少と起こり得る絶滅は、遺伝的多様性における代わりのない損失となり、これらの固有の代わりのいない種と我々自身の両方に利益をもたらすであろう、緊急を要する世界的な保全の努力に対する動機となっている。

【０３４７】

［データ生成およびアラインメント］

アプリケーションの中の座標は、複数の配列アラインメントを使用してhg19にマッピングされる他の種におけるバリアントに対する座標を含む、ヒトゲノムbuild UCSC hg19/GRCh37を参照する。タンパク質コーディングDNA配列に対する正規の転写産物および99種の脊椎動物ゲノムの複数の配列アラインメントおよび枝長が、UCSCゲノムブラウザからダウンロードされた。

【０３４８】

Exome Aggregation Consortium(ExAC)/Genome Aggregation Database(gnomAD exomes)v2.0から、ヒトエクソン多型データを取得した。24体のチンパンジー、13体のボノボ、27体のゴリラ、および10体のオランウータンに対する全体のゲノムシーケンシングデータおよび遺伝子型を備える、大型類人猿ゲノムシーケンシングプロジェクトからの霊長類変異データを取得した。チンパンジーおよびボノボの別の研究からの35体のチンパンジーからの変異も含めたが、バリアントコーリング方法の違いにより、集団分析からはこれらを除外し、深層学習モデルの訓練にのみそれらを使用した。加えて、アカゲザルの個体16体およびマーモセットの個体9体が、これらの種に対する元のゲノムプロジェクトにおける変異を評価するために使用されたが、個体レベルの情報が利用可能ではなかった。アカゲザル、マーモセット、ブタ、ウシ、ヤギ、ネズミ、ニワトリ、およびゼブラフィッシュについての変異データをdbSNPから取得した。dbSNPは追加のオランウータンバリアントも含んでおり、我々はそれを深層学習モデルの訓練にのみ使用した。それは、個体の遺伝子型情報が集団分析に利用可能ではなかったからである。平衡選択による影響を避けるために、集団分析のための延長された主要組織適合遺伝子複合体領域(chr6:28,477,797-33,448,354)内からのバリアントも除外した。

【０３４９】

ヒトタンパク質コーディング領域へのオーソロガスな1対1のマッピングを確実にし、偽遺伝子へのマッピングを防ぐために、99種の脊椎動物の多種アラインメントを使用した。バリアントが基準方向/代替方向のいずれかで発生した場合、バリアントを同一状態であるものとして受け入れた。バリアントがヒトと他の種の両方において同じ予測されるタンパク質コーディング結果を有することを確実にするために、ミスセンスバリアントと同義バリアントの両方に対して、コドンの中の他の2つのヌクレオチドが種間で同一であることを要求した。分析に含まれる各種からの多型は補足データファイル1において列挙され、詳細な尺度は補足テーブル1に示されている。

【０３５０】

4つのアレル頻度カテゴリの各々に対して(図49A)、96個の潜在的なトリヌクレオチドコンテクストの各々における同義バリアントとミスセンスバリアントの予想される数を推定するために、および、変異率を訂正するために(図51および補足テーブル7、8(図59))、イントロン領域における変異を使用した。我々はまた、同一状態のCpGジヌクレオチドおよび非CpGジヌクレオチドバリアントを別々に分析し、ミスセンス/同義比が両方のクラスに対してアレル頻度スペクトラムにわたって平坦であったことを検証した。これは、CpGバ

リアントと非CpGバリアントの両方に対して、それらの変異率の違いが大きいにもかかわらず、我々の分析が適用できることを示している(図52A、図52B、図52C、および図52D)。

【０３５１】

［他の種における多型と同一状態であるヒトミスセンスバリアントの枯渇率］

他の種に存在するバリアントがヒトにおいて一般的なアレル頻度(>0.1%)で耐えられるかどうかを評価するために、他の種における変異と同一状態であったヒトバリアントを特定した。バリアントの各々に対して、それらをヒト集団におけるそれらのアレル頻度に基づいて、4つのカテゴリ(シングルトン、シングルトンより多い～0.01%、0.01%～0.1%、>0.1%)のうちの1つに割り当て、稀(<0.1%)なバリアントと一般的(>0.1%)なバリアントとの間でのミスセンス/同義比(MSR)の低下を推定した。一般的なヒトアレル頻度(>0.1%)での同一状態のミスセンスバリアントの枯渇率は、ヒトにおける一般的なアレル頻度で自然選択により除去されるのに十分に有害な他の種からのバリアントの割合を示す。

【０３５２】

【数４０】

$$\%\,\mathrm{depletion} = \frac{\mathrm{MSR}_{\mathrm{rare}} - \mathrm{MSR}_{\mathrm{comm}}}{\mathrm{MSR}_{\mathrm{rare}}}$$

【０３５３】

ミスセンス/同義比と枯渇の割合は、種ごとに計算され、図50Bおよび補足テーブル2に示される。加えて、チンパンジーの一般的なバリアント(図49B)、チンパンジーのシングルトンバリアント(図49C)、および哺乳類バリアント(図50A)について、稀なバリアントと一般的なバリアントとの間でのミスセンス/同義比の差が有意であったかどうかを検定するために、2×2の分割表上で相同性のカイ二乗検定(χ^2)を実行した。

【０３５４】

シーケンシングは大型類人猿ゲノムプロジェクトからの限られた数の個体についてのみ実行されたので、一般的なチンパンジー集団において稀(<0.1%)または一般的(>0.1%)であったサンプリングされたバリアントの割合を推定するために、ExACからのヒトアレル頻度スペクトラムを使用した。ExACアレル頻度に基づいて24人のヒトのコホートをサンプリングし、このコホートにおいて一度、または一度より多く観察されたミスセンスバリアントを特定した。一度より多く観察されたバリアントは一般の集団において一般的(>0.1%)である確率が99.8%であったが、コホートにおいて一度しか観察されなかったバリアントは一般の集団において一般的である確率が69%であった。より遠縁の哺乳類におけるミスセンスバリアントに対する観察される枯渇が、よく保存されておりしたがってより正確にアラインメントされている遺伝子の区別できない影響によるものではなかったことを検証するために、ヒトと比較した11種の霊長類および50種の哺乳類の複数の配列アラインメントにおいて50%を超える平均ヌクレオチド相同性を持つ遺伝子のみに限定して、上記の分析を繰り返した(補足テーブル3参照)。

【０３５５】

これは、結果に大きな影響を与えることなく、分析からヒトタンパク質コーディング遺伝子の約7%を取り除いた。加えて、結果がバリアントコーリングまたは家畜化によるアーティファクト(dbSNPから選択された種の大半が家畜化されているので)の問題により影響を受けなかったことを確実にするために、種間多型の代わりに、近縁の種のペアからの固定された置換を使用して分析を繰り返した(図50D、補足テーブル4、および補足データファイル2)。

【０３５６】

［ヒト、霊長類、哺乳類、および他の脊椎動物に対する多型データのClinVar分析］

他の種と同一状態であるバリアントの臨床上の影響を調査するために、矛盾する病原性のアノテーションを持っていたバリアントまたは有意性が不確かなバリアントとしてのみ

ラベリングされたバリアントを除いて、ClinVarデータベースをダウンロードした。補足テーブル9に示されるフィルタリングステップの後で、合計で、病原性カテゴリの中の24853個のミスセンスバリアントおよび良性カテゴリの中の17775個のミスセンスバリアントがある。

【０３５７】

ヒト、ヒト以外の霊長類、哺乳類、および他の脊椎動物における変異と同一状態であった、病原性ClinVarバリアントおよび良性ClinVarバリアントの数をカウントした。ヒトについては、ExACアレル頻度からサンプリングされた30人のヒトのコホートをシミュレートした。各種に対する良性バリアントと病原性バリアントの数が補足テーブル10に示されている。

【０３５８】

［モデル訓練のための良性バリアントとラベリングされていないバリアントの生成］

機械学習のために、ヒトおよびヒト以外の霊長類からの大部分が一般的である良性ミスセンスバリアントの良性訓練データセットを構築した。このデータセットは、一般的なヒトバリアント(>0.1%のアレル頻度、83546個のバリアント)、ならびにチンパンジー、ボノボ、ゴリラ、およびオランウータン、アカゲザル、およびマーモセットからのバリアント(301690個の固有の霊長類バリアント)を備える。各源が寄与する良性訓練バリアントの数が補足テーブル5に示されている。

【０３５９】

トリヌクレオチドコンテクスト、シーケンシングカバレッジ、およびそれらの種とヒトとの間のアラインメント可能性を考慮するために、照合されたラベリングされた良性バリアントのセットとバリアントのラベリングされていないセットとを区別するように、深層学習ネットワークを訓練した。ラベリングされていない訓練データセットを得るために、正規のコーディング領域におけるすべての潜在的なミスセンスバリアントで開始した。ExACからの123136個のエクソンにおいて観察されたバリアントと、開始コドンまたは終止コドンにおけるバリアントとを除外した。合計で、68,258,623個のラベリングされていないミスセンスバリアントが生成された。これは、シーケンシングカバレッジが悪い領域、および霊長類バリアントに対する照合されたラベリングされていないバリアントを選択するときにヒトゲノムと霊長類ゲノムとの間で1対1のアラインメントがなかった領域について修正するために、フィルタリングされた。

【０３６０】

ラベリングされた良性バリアントの同じセットと、ラベリングされていないバリアントの8個のランダムにサンプリングされたセットとを使用する、8個のモデルを訓練し、それらの予測の平均をとることによって、コンセンサス予測を得た。妥当性確認および検定のために、10000個の霊長類バリアントの2つのランダムにサンプリングされたセットも除外し、それらについては訓練を保留した(補足データファイル3)。これらのセットの各々に対して、トリヌクレオチドコンテクストによって照合された10000個のラベリングされていないバリアントをサンプリングし、これらを、異なる分類アルゴリズム間で比較するときに各分類器の閾値を正規化するために使用した(補足データファイル4)。他の実装形態では、2個～500個にわたる、より少数のまたは追加のモデルがアンサンブルにおいて使用され得る。

【０３６１】

一方は一般的なヒトバリアントのみを用いて訓練され、もう一方は一般的なヒトバリアントと霊長類バリアントの両方を含む完全な良性のラベリングされたデータセットを用いた訓練された、深層学習ネットワークの2つのバージョンの分類の正確さを評価した。

【０３６２】

［深層学習ネットワークのアーキテクチャ］

各バリアントに対して、病原性予測ネットワークは、対象のバリアントを中心とする長さ51のアミノ酸配列と、二次構造および溶媒接触性ネットワーク(図2および図3)の出力とを、中心の場所において置換されるミスセンスバリアントとともに入力として取り込む。

11種の霊長類のための1つの場所頻度行列と、霊長類を除く50種の哺乳類のための1つの場所頻度行列と、霊長類と哺乳類を除く38種の脊椎動物のための1つの場所頻度行列とを含む、3つの長さ51の場所頻度行列が、99種の脊椎動物の複数の配列アラインメントから生成される。

【０３６３】

二次構造深層学習ネットワークは、各アミノ酸の場所における3状態の二次構造、すなわちαヘリックス(H)、βシート(B)、およびコイル(C)を予測する(補足テーブル11)。溶媒接触性ネットワークは、各アミノ酸の場所における3状態の溶媒接触性、すなわち、埋もれている(buried)(B)、中間(intermediate)(I)、および露出している(exposed)(E)を予測する(補足テーブル12)。両方のネットワークが、入力としてフランキングアミノ酸配列のみを取り込み、Protein DataBankにおける既知の冗長ではない結晶構造からのラベルを使用して訓練された(補足テーブル13)。事前訓練された3状態二次構造ネットワークおよび3状態溶媒接触性ネットワークへの入力のために、やはり長さが51であり深さが20である、すべての99種の脊椎動物に対する複数の配列アラインメントから生成された単一の長さ場所頻度行列を使用した。Protein DataBankからの既知の結晶構造についてネットワークを事前訓練した後で、二次構造および溶媒モデルに対する最終的な2つの層が除去され、ネットワークの出力は病原性モデルの入力に直接接続された。3状態2次構造予測モデルについて達成される最良の検定の正確さは79.86%であった(補足テーブル14)。結晶構造を有していた約4000個のヒトタンパク質に対するDSSP(Define Secondary Structure of Proteins)とアノテートされた構造ラベルを使用するときと、予測される構造ラベルのみを使用するときとでニューラルネットワークの予測を比較すると、大きな差はなかった(補足テーブル15)。

【０３６４】

病原性予測のための我々の深層学習ネットワーク(PrimateAI)と、二次構造および溶媒接触性を予測するための深層学習ネットワークの両方が、残基ブロックのアーキテクチャを採用した。PrimateAIの詳細なアーキテクチャは、(図3)および補足テーブル16(図4A、図4B、および図4C)において説明されている。二次構造および溶媒接触性を予測するためのネットワークの詳細なアーキテクチャは、図6および補足テーブル11(図7Aおよび図7B)および12(図8Aおよび図8B)において説明されている。

【０３６５】

［10000個の霊長類バリアントの保留された検定セットに対する分類器性能のベンチマーキング］

深層学習ネットワーク、ならびに、データベースdbNSFPから予測スコアを取得した他の20個のこれまでに公開されている分類器のベンチマークをとるために、検定データセットにおいて10000個の保留された霊長類バリアントを使用した。10000個の保留された霊長類バリアント検定セットに対する分類器の各々の性能も図28Aにおいて与えられる。異なる分類器は大きく変動するスコア分布を有していたので、各分類器に対する50パーセンタイル閾値を特定するために、トリヌクレオチドコンテクストによって検定セットと照合された10000個のランダムに選択されたラベリングされていないバリアントを使用した。方法間での公平な比較を確実にするために、その分類器に対して50パーセンタイルの閾値で良性であるものとして分類された、10000個の保留された霊長類バリアント検定セットの中のバリアントの割合について、各分類器のベンチマークをとった。

【０３６６】

分類器の各々に対して、50パーセンタイル閾値を使用して良性であるものとして予測される保留された霊長類検定バリアントの割合も、図28Aおよび補足テーブル17(図34)に示されている。PrimateAIの性能は、バリアントの場所におけるアラインメントされた種の数に関してロバストであり、哺乳類からの十分な保存情報が利用可能である限り全般的に良好な性能であることも示し、これは大半のタンパク質コーディング配列について当てはまる(図57)。

【０３６７】

［DDD研究からのde novoバリアントの分析］

DDD研究からの公開されているde novoバリアントと、SSC自閉症研究における健康な兄弟の対照群からのde novoバリアントとを取得した。DDD研究はde novoバリアントの信頼性レベルを提供しており、我々は、バリアントコーリングエラーによる潜在的な偽陽性として、閾値が0.1未満であるバリアントをDDDデータセットから除外した。一実装形態では、全体で、DDDの影響を受けている個人から3512個のミスセンスde novoバリアントと、健康な対照群からの1208個のミスセンスde novoバリアントがあった。99種の脊椎動物の複数配列アラインメントのためにUCSCによって使用された正規の転写産物アノテーションは、DDDにより使用される転写産物アノテーションとわずかに異なり、ミスセンスバリアントの総数の小さな違いをもたらしている。DDDの影響を受けている個人におけるde novoミスセンスバリアントと、自閉症研究からの影響を受けていない兄弟の対照群におけるde novoミスセンスバリアントとを、この分類方法が区別する能力について評価した。各分類器に対して、2つの分布に対する予測スコア間の差のウィルコクソンの順位和検定からのP値を報告した(補足テーブル17(図34))。

【０３６８】

同じ疾患遺伝子内での良性変異と病原性変異を様々な分類器が区別する際の正確さを測るために、DDDコホートにおけるde novoタンパク質切断変異についてエンリッチされた(P<0.05、ポワソン正確検定)、605個の遺伝子のサブセットに対して分析を繰り返した(補足テーブル18)。これらの605個の遺伝子内で、予想を超えるde novoミスセンス変異の3/1エンリッチメントに基づき、DDDデータセットの中のde novoバリアントの3分の2が病原性であり、3分の1が良性であったと推定した。最小限の不完全な浸透と、健康な対照群におけるde novoミスセンス変異が良性であったこととを仮定した。各分類器に対して、同じ数の良性の予測または病原性の予測を生み出した閾値を、これらのデータセットにおいて観察される経験的な割合として特定し、この閾値を、症例群vs対照群におけるde novo変異を各分類器が区別する際の正確さを推定するためのバイナリカットオフとして使用した。受信者動作特性曲線を構築するために、de novo DDDバリアントの病原性の分類を真陽性のコールとして扱い、健康な対照群における病原性としてのde novoバリアントの分類を偽陽性のコールとして扱った。DDDデータセットは3分の1の良性のde novoバリアントを含むので、理論的に完璧な分類器に対する曲線下面積(AUC)は1より小さい。したがって、良性バリアントと病原性バリアントを完璧に分離する分類器は、DDD患者におけるde novoバリアントの67%を真陽性として、DDD患者におけるde novoバリアントの33%を偽陽性として、対照群におけるde novoバリアントの100%を真陰性として分類し、0.837という最大の可能なAUCを生む(図29Aおよび図29Bおよび補足テーブル19(図35))。

【０３６９】

［新しい遺伝子候補の発見］

観察されるde novo変異の数をヌル変異モデルのもとで予想される数と比較することによって、遺伝子におけるde novo変異のエンリッチメントを検定した。DDD研究において実行されるエンリッチメント分析を繰り返し、PrimateAIスコアが0.803を超えるde novoミスセンス変異のみをカウントするときに新たにゲノムワイド有意である遺伝子を報告した。0.803を超えるPrimateAI閾値を満たすミスセンスバリアントの割合(ゲノム全体で概ねすべての潜在的なミスセンス変異の5分の1)によって、de novoの損害を与えるミスセンス変異に対するゲノムワイド期待値を調整した。DDD研究ごとに、各遺伝子は4つの検定を必要とし、1つはタンパク質切断エンリッチメントを検定し、1つはタンパク質を変化させるde novo変異のエンリッチメントを検定し、両方が、DDDコホートだけのために、および神経発達トリオシーケンシングコホートのより大きなメタ分析のために検定される。タンパク質を変化させるdee novo変異のエンリッチメントは、コーディング配列内のミスセンスde novo変異のクラスタリングの検定と、Fisherの方法によって組み合わされた(補足テーブル20、21)。各遺伝子に対するP値が4つの検定の最小値から取られ、ゲノムワイド有意性が$P<6.757\times10^{-7}$として決定された(α=0.05、4つの検定を用いた18500個の遺伝子)。

【０３７０】

［ClinVar分類の正確さ］

既存の分類器の大半は、ClinVar上で訓練される分類器からの予測スコアを使用するなどして、ClinVarコンテンツ上で直接または間接的にのいずれかで訓練されるので、2017年以降に追加されたClinVarバリアントのみを使用するように、ClinVarデータセットの分析を限定した。最近のClinVarバリアントと他のデータベースとの間にはかなりの重複があったので、ExACにおいて一般的なアレル頻度(>0.1%)で見つかるバリアント、または、HGMD(Human Gene Mutation Database)、LOVD(Leiden Open Variation Database)、またはUniprot(Universal Protein Resource)に存在するバリアントを除去するために、さらにフィルタリングを行った。有意性が不確かであるものとしてだけアノテートされたバリアントおよび矛盾するアノテーションを伴うバリアントを取り除いた後で、良性のアノテーションを伴う177個のミスセンスバリアントおよび病原性のアノテーションを伴う969個のミスセンスバリアントが残った。これらのClinVarバリアントを、深層学習ネットワークと他の分類方法の両方を使用してスコアリングした。各分類器に対して、同じ数の良性予測と病原性予測を生み出した閾値を、これらのデータベースにおいて観察される経験的な割合として特定し、この閾値を、各分類器の正確さを推定するためのバイナリカットオフとして使用した(図31Aおよび図31B)。

【０３７１】

[訓練データサイズを大きくすることおよび訓練データの異なる源を使用することの影響]

深層学習ネットワークの性能に対する訓練データサイズの影響を評価するために、385236個の霊長類および一般的なヒトのバリアントの良性とラベリングされた訓練セットから、バリアントのサブセットをランダムにサンプリングし、背後の深層学習ネットワークアーキテクチャを同一に保った。各々の個別の霊長類の種からのバリアントが分類の正確さに寄与する一方で、各々の個別の哺乳類の種からのバリアントはより低い分類の正確さに寄与することを示すために、一実装形態に従って、83546個のヒトバリアントと、各種に対するランダムに選択された一定の数のバリアントとを備える訓練データセットを使用して、深層学習ネットワークを訓練し、背後のネットワークアーキテクチャを再び同じに保った。訓練セットに追加したバリアントの一定の数(23380)は、ミスセンスバリアントの数が最小である種、すなわちボノボにおいて利用可能なバリアントの総数であった。各分類器に対する性能の中央値を得るために、訓練手順を5回繰り返した。

【０３７２】

［シーケンシングされる霊長類集団の数の増大に伴うすべての潜在的なヒトミスセンス変異の飽和］

ExACにおいて観察される一般的なヒトミスセンスバリアント(>0.1%のアレル頻度)のトリヌクレオチドコンテクストに基づいてバリアントをシミュレートすることによって、504種の現存する霊長類の種において存在する一般的なバリアントによる、すべての約7000万個の潜在的なヒトミスセンス変異の予想される飽和を調査した。各霊長類の種に対して、ヒトにおいて観察される一般的なミスセンスバリアントの数(アレル頻度が0.1%を超える約83500個のミスセンスバリアント)の4倍をシミュレートした。それは、ヒトが、他の霊長類の種と比べて個体あたりのバリアントの数が概ね半分であり、0.1%を超えるアレル頻度では、純化選択によって約50%のヒトミスセンスバリアントが取り除かれているからである(図49A)。

【０３７３】

調査されるヒトのコホートのサイズの増大に伴って発見される一般的なヒトミスセンスバリアント(>0.1%のアレル頻度)の割合をモデル化するために(図56)、ExACアレル頻度に従って遺伝子型をサンプリングし、これらのシミュレートされるコホートにおいて少なくとも1回観察された一般的なバリアントの割合を報告した。

【０３７４】

一実装形態では、PrimateAIスコアの現実的な応用のために、優性遺伝モードを伴う遺伝子においては、対照群と比較した症例群におけるde novoバリアントのエンリッチメントに基づいて(図21D)、>0.8という閾値が病原性の可能性が高いという分類に対して好ま

しく、<0.6が良性である可能性が高いという分類に対するものであり、0.6～0.8が中間であり、劣性遺伝モードを伴う遺伝子においては、>0.7という閾値が病原性である可能性が高いという分類に対するものであり、<0.5が良性である可能性が高いという分類に対するものである。

【０３７５】

図2は、本明細書で「PrimateAI」と呼ばれる、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2において、1Dは1次元畳み込み層を指す。予測される病原性は、0(良性)から1(病原性)までの目盛り上にある。ネットワークは、ヒトアミノ酸(AA)基準およびバリアントを中心とする代替配列(51個のAA)、99種の脊椎動物の種から計算された位置特定的重み行列(PWM)保存プロファイル、ならびに二次構造および溶媒接触性予測深層学習ネットワークの出力を入力として取り込み、この深層学習ネットワークは、3状態のタンパク質二次構造(ヘリックス-H、βシート-B、およびコイル-C)と、3状態の溶媒接触性(埋もれている-B、中間-I、および露出している-E)とを予測する。

【０３７６】

図3は、病原性分類のための深層学習ネットワークアーキテクチャであるPrimateAIの概略図を示す。モデルへの入力は、基準配列と置換されるバリアントを伴う配列との両方に対するフランキング配列の51個のアミノ酸(AA)と、霊長類、哺乳類、および脊椎動物のアラインメントからの3つの長さ51AAの位置特定的重み行列により表される保存率と、事前訓練された二次構造ネットワークおよび溶媒接触性ネットワークの出力(やはり長さは51AAである)とを含む。

【０３７７】

図4A、図4B、および図4Cは、病原性予測深層学習モデルPrimateAIの例示的なモデルアーキテクチャの詳細を示す、補足テーブル16である。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。モデルへの入力は、バリアントの周りのフランキングアミノ酸配列に対する位置特定的頻度行列(長さ51AA、深さ20)、ワンホット符号化された(one-hot encoded)ヒト基準配列および代替配列(長さ51AA、深さ20)、ならびに、二次構造および溶媒接触性モデルからの出力(長さ51AA、深さ40)である。

【０３７８】

示される例は1D畳み込みを使用する。他の実装形態では、このモデルは、2D畳み込み、3D畳み込み、拡張畳み込みまたは膨張畳み込み、転置畳み込み(transposed convolution)、分離可能畳み込み(separable convolution)、および深さごとの(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。一部の層は、シグモイドまたは双曲線正接などの飽和する非線形性と比較して確率的勾配降下の収束を大きく加速する、ReLU活性化関数も使用する。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、parametric ReLU、leaky ReLU、および指数関数的線形ユニット(ELU)がある。

【０３７９】

一部の層はバッチ正規化(IoffeおよびSzegedy、2015年)も使用する。バッチ正規化に関して、畳み込みニューラルネットワーク(CNN)における各層の分布は訓練の間に変化し、層によって変化する。このことは、最適化アルゴリズムの収束速度を低下させる。バッチ正規化はこの問題を克服するための技法である。バッチ正規化層の入力をxで表し、その出力をzを使用して表すと、バッチ正規化はxについての以下の変換を適用する。

【０３８０】

【数４１】

$$z = \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} \gamma + \beta$$

【０３８１】
バッチ正規化は、μおよびσを使用して入力xに対する平均-分散の正規化を適用し、γおよびβを使用してそれを線形にスケーリングしてシフトする。正規化パラメータμおよびσは、指数移動平均と呼ばれる方法を使用して訓練セットにわたって現在の層に対して計算される。言い換えると、それらは訓練可能なパラメータではない。対照的に、γおよびβは訓練可能なパラメータである。訓練の間に計算されるμおよびσの値は、推論の間にフォワードパスにおいて使用される。
【０３８２】
図5および図6は、タンパク質の二次構造および溶媒接触性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。モデルに対する入力は、RaptorXソフトウェア(Protein Data Bank配列について訓練するための)または99種の脊椎動物のアラインメント(ヒトタンパク質配列についての訓練および推論のための)によって生成される保存率を使用した、位置特定的重み行列である。長さが51AAである、第2の層から最後の層の出力は、病原性分類のための深層学習ネットワークに対する入力になる。
【０３８３】
図7Aおよび図7Bは、3状態二次構造予測深層学習(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補足テーブル11である。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。モデルへの入力は、バリアントの周りのフランキングアミノ酸配列に対する位置特定的頻度行列(長さ51AA、深さ20)であった。
【０３８４】
図8Aおよび図8Bは、3状態溶媒接触性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補足テーブル12である。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。モデルへの入力は、バリアントの周りのフランキングアミノ酸配列に対する位置特定的頻度行列(長さ51AA、深さ20)であった。
【０３８５】
図20は、重要な機能ドメインに対してアノテートされた、SCN2A遺伝子における各アミノ酸の位置における予測される病原性スコアを示す。遺伝子に沿ってプロットされているのは、各アミノ酸の位置におけるミスセンス置換に対する平均PrimateAIスコアである。
【０３８６】
図21Dは、訓練を保留された10000個の一般的な霊長類バリアントの検定セットに対する良性の結果を予測することにおける分類器の比較を示す。y軸は、各分類器の閾値を変異率について照合された10000個のランダムなバリアントのセットについての50パーセンタイルスコアへと正規化した後の、良性であるものとして正しく分類された霊長類バリアントの百分率を表す。
【０３８７】
図21Eは、Deciphering Developmental Disorders(DDD)患者において発生するde novoミスセンスバリアントに対するPrimateAI予測スコアの分布を、影響を受けていない兄弟と比較して、対応するウィルコクソンの順位和のP値とともに示す。
【０３８８】
図21Fは、DDD症例群vs対照群におけるde novoミスセンスバリアントを分離することにおける分類器の比較を示す。ウィルコクソンの順位和検定のP値が各分類器に対して示されている。
【０３８９】
図22A、図22B、図22C、図22D、および図22Eは、P<0.05である605個のDDD遺伝子内の分類の正確さを示す。図22Aは、de novoタンパク質切断変異(P<0.05)に対して有意であった605個の関連する遺伝子内の、DDDコホートからの影響を受けている個人における予想を超えるde novoミスセンス変異のエンリッチメントを示す。図22Bは、605個の関連する遺伝子内での、DDD患者vs影響を受けていない兄弟において発生するde novoミスセンスバリア

ントに対するPrimateAI予測スコアの分布を、対応するウィルコクソンの順位和のP値とともに示す。

【０３９０】

図22Cは、605個の遺伝子内での、症例群vs対照群におけるde novoミスセンスバリアントを分離する際の様々な分類器の比較を示す。y軸は、各分類器に対するウィルコクソンの順位和検定のP値を示す。

【０３９１】

図22Dは、受信者動作特性曲線上で示される、様々な分類器の比較を、各分類器に対して示されるAUCとともに示す。

【０３９２】

図22Eは、各分類器に対する分類の正確さとAUCを示す。示される分類の正確さは、図22Aにおいて示されるエンリッチメントに基づいて予測されるのと同じ数の病原性バリアントと良性バリアントを分類器が予測するような閾値を使用した、真陽性と真陰性のエラー率の平均である。DDD de novoミスセンスバリアントの33%がバックグラウンドとなるという事実を考慮するために、完璧な分類器に対する最大の達成可能なAUCが点線で示される。

【０３９３】

図23A、図23B、図23C、および図23Dは、訓練のために使用されるデータの、分類の正確さに対する影響を示す。深層学習ネットワークは、完全なデータセット(385236個のバリアント)まで、増大する数の霊長類およびヒトの一般的なバリアントを用いて訓練される。図23Aにおいて、ネットワークの各々の分類性能のベンチマークが、10000個の保留された霊長類バリアントと、DDD症例群vs対照群におけるde novoバリアントとに対して正確さについてとられる。

【０３９４】

図23Bおよび図23Cは、一実装形態による、83546個の一般的なヒトバリアントと、単一の霊長類の種または哺乳類の種からの23380個のバリアントとを備えるデータセットを使用して訓練された、ネットワークの性能を示す。10000個の保留された霊長類バリアント(図23B)と、DDD症例群vs対照群におけるde novoミスセンスバリアント(図23C)とについてベンチマークがとられた、一般的な変異の異なる源を用いて訓練された各ネットワークに対して、結果が示されている。

【０３９５】

図23Dは、504種の現存する霊長類の種における同一状態の一般的なバリアント(>0.1%)による、すべての潜在的な良性のヒトミスセンスの位置の予想される飽和を示す。y軸は少なくとも1つの霊長類の種において観察されるヒトミスセンスバリアントの割合を示し、CpGミスセンスバリアントが緑で示され、すべてのミスセンスバリアントが青で示されている。各霊長類の種における一般的なバリアントをシミュレートするために、置き換えを伴うすべての潜在的な一塩基置換のセットからサンプリングし、ExACにおける一般的なヒトバリアント(>0.1%のアレル頻度)について観察されるトリヌクレオチドコンテクスト分布を照合した。

【０３９６】

図24は、一般的な霊長類バリアントの確認に対するシーケンシングカバレッジの影響を訂正することを示す。ヒト以外の霊長類の種において所与のバリアントを観察する確率は、ExAC/gnomADエクソンデータセットの中のその位置におけるシーケンシング深度と逆の相関がある。対照的に、より低いgnomADリードの深さは、その位置における一般的なヒトバリアント(>0.1%のアレル頻度)を観察する確率に影響を与えなかった。それは、シーケンシングされる多数のヒトエクソンが、一般的な変異の確認をほとんど保証されたものにするからである。ネットワークを訓練するための霊長類バリアントの各々に対する一致したバリアントを選ぶとき、あるバリアントを選ぶ確率は、変異率および遺伝子変換を考慮するためのトリヌクレオチドコンテクストに対する照合に加えて、シーケンシングの深さの影響について調整された。

【０３９７】
図25A、図25B、図25C、および図26は、開示されたニューラルネットワークによるタンパク質モチーフの認識を示す。図25A、図25B、および図25Cに関して、タンパク質ドメインのニューラルネットワークによる認識を例示するために、3つの異なるタンパク質ドメインの中の各アミノ酸位置におけるバリアントに対する平均PrimateAIスコアを示す。図25Aにおいて、反復するGXXモチーフの中にグリシンを伴う、COL1A2というコラーゲン鎖が強調されている。コラーゲン遺伝子における臨床的に特定されている変異は、大部分がGXXリピートにおけるグリシンのミスセンス変異によるものであり、それは、これらがコラーゲンの正常な組み立てと干渉し、強いドミナントネガティブ効果を及ぼすからである。図25Bにおいて、IDSスルファターゼ酵素の活性サイトが強調されており、これは、ホルミルグリシンへと翻訳後修飾されるシステインを活性サイトに含む。図25Cにおいて、MYC転写因子のbHLHzipドメインが示されている。基本ドメインは、負に荷電した糖リン酸の骨格と相互作用する正に荷電したアルギニンおよびリジン残基(強調されている)を介して、DNAに接触する。ロイシンジッパードメインは、二量体化のために決定的に重要である、7個のアミノ酸だけ離隔されたロイシン残基(強調されている)を備える。

【０３９８】
図26は、バリアントに対する予測される深層学習スコアへの、バリアントの中および周りの各位置を摂動させることの影響を示す線のプロットを含む。バリアントの周りの近くのアミノ酸(位置-25～+25)における入力をシステム的に0に設定し、ニューラルネットワークによるバリアントの予測される病原性の変化を測定した。プロットは、5000個のランダムに選択されたバリアントに対する各々の近くのアミノ酸位置における摂動に対する、予測される病原性スコアの平均の変化を示す。

【０３９９】
図27は、重みの相関パターンがBLOSUM62スコア行列およびGranthamスコア行列を模倣することを示す。二次構造深層学習ネットワークの最初の3つの層からの重みの相関パターンは、BLOSUM62スコア行列およびGranthamスコア行列に類似するアミノ酸間の相関を示す。左のヒートマップは、ワンホット表現を使用して符号化されたアミノ酸間の二次構造深層学習ネットワークの2つの初期アップサンプリング層に続く、第1の畳み込み層からのパラメータ重みの相関を示す。中間のヒートマップは、アミノ酸のペア間のBLOSUM62スコアを示す。右のヒートマップはアミノ酸間のGrantham距離を示す。深層学習重みとBLOSUM62スコアとの間のPearson相関は0.63($P=3.55\times10^{-9}$)である。深層学習重みとGranthamスコアとの間の相関は-0.59($P=4.36\times10^{-8}$)である。BLOSUM62スコアとGranthamスコアとの間の相関は-0.72($P=8.09\times10^{-13}$)である。

【０４００】
図28A、図28B、および図28Cは、深層学習ネットワークPrimateAIと他の分類器の性能評価を示す。図28Aは、訓練を保留された10000個の霊長類バリアントの検定セットに対する良性の結果を予測することにおける深層学習ネットワークPrimateAIの正確さと、SIFT、PolyPhen-2、CADD、REVEL、M-CAP、LRT、MutationTaster、MutationAssessor、FATHMM、PROVEAN、VEST3、MetaSVM、MetaLR、MutPred、DANN、FATHMM-MKL_coding、Eigen、GenoCanyon、integrated_fitCons、およびGERPを含む、他の分類器との比較とを示す。y軸は、変異率および遺伝子変換を考慮するためにトリヌクレオチドコンテクストについて霊長類バリアントと照合された、10000個のランダムに選択されたバリアントのセットを使用して、各分類器に対する閾値を50パーセンタイルスコアへと正規化したことに基づいて、良性であるものとして分類された霊長類バリアントの百分率を表す。

【０４０１】
図28Bは、上で列挙された20個の既存の方法とともに、DDD症例群vs対照群におけるde novoミスセンスバリアントを分離する際のPrimateAIネットワークの性能の比較を示す。y軸は、各分類器に対するウィルコクソンの順位和検定のP値を示す。

【０４０２】
図28Cは、上で列挙された20個の方法とともに、605個の疾患関連遺伝子内での、DDD症

例群vs影響を受けていない対照群におけるde novoミスセンスバリアントを分離する際のPrimateAIネットワークの性能の比較を示す。y軸は、各分類器に対するウィルコクソンの順位和検定のP値を示す。

【０４０３】

図29Aおよび図29Bは、4つの分類器の予測スコアの分布を示す。DDD症例群vs影響を受けていない対照群において発生するde novoミスセンスバリアントに対する、SIFT、PolyPhen-2、CADD、およびREVELを含む4つの分類器の予測スコアのヒストグラムを、対応するウィルコクソンの順位和のP値とともに示す。

【０４０４】

図30A、図30B、および図30Cは、605個の疾患関連遺伝子における病原性バリアントと良性バリアントを分類する際の、PrimateAIネットワークと他の分類器の正確さを比較する。図30Aの散布図は、DDD症例群vs対照群に対する分類器の各々の性能(y軸)と、保留された霊長類データセットに対する良性予測の正確さ(x軸)とを示す。図30Bは、各分類器に対して示される曲線下面積(AUC)とともに、受信者動作特性(ROC)曲線上に示される、605個の遺伝子内での症例群vs対照群におけるde novoミスセンスバリアントを分離することにおいて異なる分類器を比較する。図30Cは、図28A、図28B、および図28Cにおいて列挙される、PrimateAIネットワークおよび20個の分類器に対する分類の正確さとAUCを示す。示される分類の正確さは、図22Aにおけるエンリッチメントに基づいて予想されるものと同じ数の病原性バリアントと良性バリアントを分類器が予測するような閾値を使用した、真陽性と真陰性の率の平均である。DDD症例群におけるde novoミスセンスバリアントは67%が病原性バリアントであり33%が良性であり、対照群におけるde novoミスセンスバリアントは100%良性であると仮定して、完璧な分類器に対する最大の達成可能なAUCが点線で示されている。

【０４０５】

図31Aおよび図31Bは、専門家により精選されたClinVarバリアントに対する分類器の性能と、経験的なデータセットに対する性能との相関を示す。図31Aの散布図は、ヒトのみのデータで訓練された、またはヒト+霊長類のデータで訓練された、20個の他の分類器の各々およびPrimateAIネットワークに対する、10000個の保留された霊長類バリアントについての分類の正確さ(x軸)と、ClinVarバリアントについての分類の正確さ(y軸)とを示す。Spearman相関係数rhoおよび関連するP値が示されている。分類器を訓練するために使用されなかったデータへと評価を限定するために、2017年1月から2017年11月の間に追加されたClinVarバリアントのみを使用し、ExAC/gnomADからの一般的なヒトバリアント(>0.1%のアレル頻度)を除外した。示されるClinVar分類の正確さは、ClinVarデータセットにおいて観察されるのと同じ数の病原性バリアントと良性バリアントを分類器が予測するような閾値を使用した、真陽性と真陰性の率の平均である。

【０４０６】

図31Bの散布図は、ヒトのみのデータで訓練された、またはヒト+霊長類のデータで訓練された、20個の他の分類器およびPrimateAIネットワークの各々に対する、DDD症例群vs対照群の完全なデータセット(x軸)と、ClinVarバリアントに対する分類の正確さ(y軸)を示す。

【０４０７】

図32は、訓練のための6367個の、妥当性確認のための400個の、および検定のための500個の関連しないタンパク質配列を使用した、Protein DataBankからのアノテートされたサンプルに対する、3状態二次構造予測モデルおよび3状態溶媒接触性予測モデルの性能を示す、補足テーブル14である。配列の相同性が25%未満であるタンパク質のみがProtein DataBankから選択された。3つのクラスは二次構造と溶媒接触性のいずれについても大きく偏ってはいないので、深層学習ネットワークの正確さを性能の尺度として報告する。

【０４０８】

図33は、予測される二次構造ラベルを使用する深層学習ネットワークとともに、DSSPデータベースからのヒトタンパク質のアノテートされた二次構造ラベルが利用であるときに

それを使用した、深層学習ネットワークの性能比較を示す補足テーブル15である。

【０４０９】

図34は、評価した20個の分類器の各々についての、10000個の保留された霊長類バリアントに対する正確さの値と、DDD症例群vs対照群におけるde novoバリアントに対するp値とを示す補足テーブル17である。ヒトのデータのみを用いたPrimateAIモデルは、一般的なヒトバリアント(集団において>0.1%である83500個のバリアント)のみからなるラベリングされた良性の訓練データセットを使用して訓練された我々の深層学習ネットワークであるが、ヒト+霊長類のデータを用いたPrimateAIモデルは、一般的なヒトバリアントと霊長類バリアントの両方を備える、385000個のラベリングされた良性バリアントの完全なセットについて訓練された我々の深層学習ネットワークである。

【０４１０】

図35は、605個の疾患関連遺伝子に制約された、DDD症例群vs対照群データセットにおけるde novoバリアントに対する異なる分類器の性能の比較を示す、補足テーブル19である。異なる方法の間で正規化するために、各分類器に対して、DDDセットと対照群セットとにおけるエンリッチメントに基づいて予想されるのと同じ数の病原性バリアントおよび良性バリアントを分類器が予測するような閾値を特定した。示される分類の正確さは、この閾値における真陽性と真陰性のエラー率の平均である。

【０４１１】

図49A、図49B、図49C、図49D、および図49Eは、ヒトアレル頻度スペクトラムにわたるミスセンス/同義比を示す。図49Aは、ExAC/gnomADデータベースからの123136人のヒトにおいて観察されるミスセンスバリアントおよび同義バリアントが、アレル頻度によって4つのカテゴリへと分割されたことを示す。灰色の影付きの棒は各カテゴリにおける同義バリアントのカウントを表し、濃緑の棒はミスセンスバリアントを表す。各棒の高さは各アレル頻度カテゴリにおける同義バリアントの数に対してスケーリングされ、ミスセンス/同義カウントおよび比は変異率を調整した後で表示される。図49Bおよび図49Cは、チンパンジーの一般的なバリアント(図49B)およびチンパンジーのシングルトンバリアント(図49C)と同一状態(IBS)であるヒトミスセンスバリアントおよび同義バリアントに対する、アレル頻度スペクトラムを示す。稀なヒトアレル頻度(<0.1%)と比較して一般的なヒトアレル頻度(>0.1%)を持つチンパンジーミスセンスバリアントの枯渇率が、付随するカイ二乗(χ^2)検定のP値とともに、赤いボックスによって示されている。

【０４１２】

図49Dは、ヒト以外の霊長類の種のうちの少なくとも1つにおいて観察されるヒトバリアントを示す。図49Eは、ClinVarデータベース全体における良性ミスセンスバリアントと病原性ミスセンスバリアントのカウント(上の行)を、ExAC/gnomADアレル頻度からサンプリングされる30人のヒトのコホートにおけるClinVarバリアントのカウント(中間の行)、および霊長類において観察されるバリアントのカウント(下の行)と比較して示す。矛盾する良性と病原性の判定と、有意性が不確かなものとしてだけアノテートされたバリアントは、除外された。

【０４１３】

図50A、図50B、図50C、および図50Dは、他の種と同一状態であるミスセンスバリアントに対する純化選択を示す。図50Aは、4種の霊長類以外の哺乳類の種(ネズミ、ブタ、ヤギ、およびウシ)において存在するバリアントと同一状態である、ヒトミスセンスバリアントおよび同義バリアントに対するアレル頻度スペクトラムを示す。一般的なヒトアレル頻度(>0.1%)を持つミスセンスバリアントの枯渇率が、付随するカイ二乗(χ^2)検定のP値とともに、赤いボックスによって示されている。

【０４１４】

図50Bは、一般的なヒトアレル頻度(>0.1%)で他の種において観察されるミスセンスバリアントの枯渇率と、枝長(ヌクレオチドの位置当たりの置換の平均の数)の単位で表される、ヒトからの種の進化的距離とを比較して示す散布図である。各種とヒトとの間の枝長の合計が、種の名前の隣に示されている。親類の個体を含んでいたゴリラを除いて、シング

ルトンバリアントおよび一般的なバリアントに対する枯渇値が、バリアント頻度が入手可能であった種に対して示されている。

【０４１５】

図50Cは、ExAC/gnomADアレル頻度からサンプリングされた30人のヒトのコホートにおける良性ミスセンスバリアントと病原性ミスセンスバリアントのカウント(上の行)を、霊長類において観察されるバリアントのカウント(中間の行)、ならびに、ネズミ、ブタ、ヤギ、およびウシにおいて観察されるバリアントのカウント(下の行)と比較して示す。矛盾する良性と病原性の判定と、有意性が不確かなものとしてだけアノテートされたバリアントは、除外された。

【０４１６】

図50Dは、一般的なヒトアレル頻度(>0.1%)で近縁の種のペアにおいて観察される固定されたミスセンス置換の枯渇率と、ヒトからの種の進化的距離(平均の枝長の単位で表される)とを比較して示す、散布図である。

【０４１７】

図51は、純化選択がないときのヒトアレル頻度スペクトラムにわたる予想されるミスセンス:同義比を示す。灰色の影付きの棒は同義バリアントの数を表し、濃緑の棒はミスセンスバリアントの数を表す。点線は同義バリアントによって形成される基準を示す。ミスセンス:同義比は各アレル頻度カテゴリに対して示される。一実装形態によれば、各アレル頻度カテゴリにおける予想されるミスセンスカウントおよび同義カウントは、123136個のエクソンを備えるExAC/gnomADデータセットからイントロンバリアントを取り込み、変異率および遺伝子変換におけるGCバイアスを考慮するバリアントのトリヌクレオチドコンテクストに基づいて、4つのアレル頻度カテゴリの各々に該当することが予想されるバリアントの割合を推定するためにそれらのイントロンバリアントを使用することによって、計算された。

【０４１８】

図52A、図52B、図52C、および図52Dは、CpGバリアントおよび非CpGバリアントに対するミスセンス:同義比を示す。図52Aおよび図52Bは、ExAC/gnomADエクソンからのすべてのバリアントを使用した、ヒトアレル頻度スペクトラムにわたるCpGバリアントに対するミスセンス:同義比(図52A)および非CpGバリアントに対するミスセンス:同義比(図52B)を示す。図52Cおよび図52Dは、一般的なチンパンジー多型と同一状態であるヒトバリアントのみに制約された、ヒトアレル頻度スペクトラムにわたるCpGバリアントに対するミスセンス:同義比(図52C)および非CpGバリアントに対するミスセンス:同義比(図52D)を示す。

【０４１９】

図53、図54、および図55は、6種の霊長類と同一状態であるヒトバリアントのミスセンス:同義比を示す。チンパンジー、ボノボ、ゴリラ、オランウータン、アカゲザル、およびマーモセットにおいて存在する変異と同一状態であるExAC/gnomADバリアントに対する、ヒトアレル頻度スペクトラムにわたるミスセンス:同義比のパターンを示す。

【０４２０】

図56は、調査されるヒトのコホートのサイズを増大させることによって発見される、新しい一般的なミスセンスバリアントの飽和を示すシミュレーションである。シミュレーションにおいて、各サンプルの遺伝子型は、gnomADアレル頻度に従ってサンプリングされた。発見された一般的なgnomADバリアントの割合は、10から100000の各サンプルサイズに対して100回のシミュレーションにわたって平均される。

【０４２１】

図57は、ゲノムにおける様々な保存プロファイルにわたるPrimateAIの正確さを示す。x軸は、99種の脊椎動物のアラインメントとの、ある配列の周りの51個のAAの百分率のアラインメント可能性を表す。y軸は、10000個の保留された霊長類バリアントの検定データセットについてベンチマークがとられた、保存ビンの各々におけるバリアントに対するPrimateAIの正確さの分類性能を表す。

【０４２２】

図58は、一般的なヒトバリアントおよびヒト以外の霊長類において存在するバリアントからのラベリングされた良性の訓練データセットへの寄与を示す、補足テーブル5である。

【０４２３】

図59は、予想されるミスセンス:同義比に対するアレル頻度の影響を示す補足テーブル8である。同義バリアントとミスセンスバリアントの予想されるカウントは、変異率および遺伝子変換バイアスを考慮するためにトリヌクレオチドコンテクストを使用して、エクソン境界から少なくとも20～30nt離れたイントロン領域におけるバリアントのアレル頻度スペクトラムに基づいて計算された。

【０４２４】

図60は、ClinVar分析を示す補足テーブル9である。一実装形態によれば、ClinVarデータベースの2017年11月のビルドからダウンロードされたバリアントが、矛盾するアノテーションを伴うミスセンスバリアントを除外し、有意性が不確かであるバリアントを削除するようにフィルタリングされ、17775個の良性バリアントおよび24853個の病原性バリアントが残った。

【０４２５】

図61は、一実装形態による、ClinVarにおいて発見された他の種からのミスセンスバリアントの数を示す補足テーブル10である。バリアントは、対応するヒトバリアントと同一状態であることと、同じコーディング結果を保証するためにリーディングフレームの中の他の2つの位置において同一のヌクレオチドを有することとが必要とされた。

【０４２６】

図62は、元のDDD研究においてはゲノムワイド有意性の閾値にこれまで達していなかった、知的障害における14個の追加の遺伝子候補の発見の一実装形態を示すテーブル1である。

【０４２７】

図63は、ClinVarにおける病原性バリアントと良性バリアントとの間のGranthamスコアの平均の差の一実装形態を示すテーブル2であり、この差は、605個の疾患関連遺伝子内でのDDD症例群vs対照群におけるde novoバリアントの差の2倍である。

【０４２８】

［データ生成］

本明細書において使用されるすべての座標は、このセクションで説明される手順を使用して複数配列アラインメントを使用してhg19にマッピングされた他の種におけるバリアントに対する座標を含めて、ヒトゲノムbuild UCSC hg19/GRCh37を参照する。ヒトとの99種の脊椎動物ゲノムのタンパク質コーディングDNA配列および複数配列アラインメントが、hg19 buildのためのUCSCゲノムブラウザからダウンロードされた(http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/alignments/knownCanonical.exonNuc.fa.gz)。複数の正規の遺伝子アノテーションを伴う遺伝子については、最長のコーディング転写産物が選択された。

【０４２９】

世界中の8つの亜集団からの123136人の個人の全エクソンシーケンシング(WES)データを収集した、Exome Aggregation Consortium(ExAC)/genome Aggregation Database(gnomAD) v2.0からヒトエクソン多型データをダウンロードした(http://gnomad.broadinstitute.org/)。ExAC VCFファイルにおいてアノテートされるようなデフォルトの品質制御フィルタを通過しないバリアント、または正規のコーディング領域の外側にあるバリアントを除外した。平衡選択による影響を避けるために、霊長類分析のための延長されたMHC領域(chr6:28,477,797～33,448,354)内からのバリアントも除外した。大型類人猿ゲノムシーケンシングプロジェクトは、24体のチンパンジー、13体のボノボ、27体のゴリラ、および10体のオランウータン(スマトラ亜種からの5体およびボルネオ亜種からの5体を含む、これらを下流分析のために折り畳んだ)に対する、全ゲノムシーケンシングデータおよび遺伝子型を提供する。チンパンジーおよびボノボについての研究は、追加の35体のチンパンジーの

ゲノム配列を提供する。しかしながら、これらの追加のチンパンジーに対するバリアントは、大型類人猿ゲノムシーケンシングプロジェクトと同じ方法を使用してコールされなかったので、それらをアレル頻度スペクトラム分析から除外し、深層学習モデルを訓練するためだけに使用した。これらの霊長類多様性研究からの変異はすでに、ヒト基準(hg19)にマッピングされていた。加えて、マーモセットおよびアカゲザルについては、16体のアカゲザルの個体および9体のマーモセットの個体がこれらの種のゲノムの元のシーケンシングにおける変異を評価するために使用されたが、個体レベルの情報は利用可能ではない。

【０４３０】

大型類人猿ゲノムシーケンシングプロジェクト4は、24体のチンパンジー、13体のボノボ、27体のゴリラ、および10体のオランウータン(スマトラ亜種からの5体およびボルネオ亜種からの5体を含む、これらを下流での分析のために折り畳んだ)に対する、全ゲノムシーケンシングデータおよび遺伝子型を提供する。チンパンジーおよびボノボについての研究は、追加の35体のチンパンジーのゲノム配列を提供する。しかしながら、これらの追加のチンパンジーに対するバリアントは、大型類人猿ゲノムシーケンシングプロジェクトと同じ方法を使用してコールされなかったので、それらをアレル頻度スペクトラム分析から除外し、深層学習モデルを訓練するためだけに使用した。これらの霊長類多様性研究からの変異はすでに、ヒト基準(hg19)にマッピングされていた。加えて、マーモセットおよびアカゲザルについては、16体のアカゲザルの個体および9体のマーモセットの個体がこれらの種のゲノムの元のシーケンシングにおける変異を評価するために使用されたが、個体レベルの情報は利用可能ではない。

【０４３１】

他の霊長類および哺乳類と比較するために、アカゲザル、マーモセット、ブタ、ウシ、ヤギ、ネズミ、ニワトリ、およびゼブラフィッシュを含む、他の種のSNPもdbSNPからダウンロードした。dbSNPは追加のオランウータンバリアントも含んでおり、それを我々は深層学習モデルを訓練するためだけに使用した。それは、個体の遺伝子型情報がアレル頻度スペクトラム分析に対して利用可能ではなかったからである。イヌ、ネコ、またはヒツジなどの他の種については、それらの種に対する限られた数のバリアントしかdbSNPが提供しないので、廃棄した。

【０４３２】

バリアントをヒトにマッピングするために、99種の脊椎動物の多種アラインメントを使用して、ヒトタンパク質コーディング領域へのオーソロガスな1:1のマッピングを確実にした。オーソロガスな多種アラインメントを使用するバリアントのマッピングが、偽遺伝子またはレトロトランスポジションを受けた配列によって引き起こされるアーティファクトを取り除くために必須であった。このアーティファクトは、多数対1のマッピングを可能にするliftOverなどのツールを使用して種間で直接SNPをマッピングするときに発生する。dbSNPにおける種のゲノムビルドが99種の脊椎動物の複数配列アラインメントにおける種のゲノムビルドと一致しなかった場合、複数配列アラインメントにおいて使用されるゲノムビルドへとバリアントを更新するためにliftOverを使用した。バリアントが基準/代替方向において発生した場合、バリアントを同一状態であるものとして受け入れた。たとえば、ヒト基準がGであり代替アレルがAであった場合、これは、基準がAであり代替アレルがGであった別の種におけるバリアントと同一状態であると見なされた。バリアントがヒトと他の種の両方において同じ予測されるタンパク質コーディング結果を有することを確実にするために、ミスセンスバリアントと同義バリアントの両方に対して、コドンの中の他の2つのヌクレオチドが種間で同一であることを要求した。分析に含まれる各種からの多型が補足データファイル1において列挙され、詳細な尺度が補足テーブル1に示されている。

【０４３３】

各dbSNP提出者バッチからのバリアントが、高品質でありヒトに正しくアラインメントされることを確実にするために、各バッチに対するミスセンス:同義比を計算し、これが2.2:1という予想される比より小さかったことと、大半の種、特に非常に大きい有効個体数

を有することが予想されるゼブラフィッシュおよびネズミが、1:1より小さい比を有していたこととを確認した。さらなる分析から、異常に高いミスセンス:同義比を有していたウシから、SNPの2つのバッチを除外した(比が1.391であるsnpBatch_1000_BULL_GENOMES_1059190.gzおよび比が2.568であるsnpBatch_COFACTOR_GENOMICS_1059634.gz)。残りのウシのバッチに対する平均のミスセンス:同義比は0.8:1であった。

【０４３４】

［ミスセンス:同義比、変異率、遺伝的浮動、およびGCバイアス(GC-Biased)遺伝子変換に対するアレル頻度の影響の訂正］

純化選択の活動に加えて、高いアレル頻度でのヒトミスセンスバリアントの観察される枯渇率は、自然選択に関連しない要因によっても影響を受け得る。集団において特定のアレル頻度で現れる自然変異の確率は、変異率、遺伝子変換、および遺伝的浮動の関数であり、これらの要因は、選択圧がなくてもアレル頻度スペクトラムにわたってミスセンス:同義比にバイアスをもたらす可能性がある。

【０４３５】

タンパク質コーディング選択がないときの各アレル頻度カテゴリにおける予想されるミスセンス:同義比を計算するために、各エクソンの31～50bp上流および21～50bp下流のイントロン領域内のバリアントを選択した。これらの領域は、延長されたスプライスモチーフの影響を避けるのに十分遠くなるように選ばれた。これらの領域は、ExAC/gnomADエクソンに対するエクソン捕捉配列の端に近いので、バリアントの公平な確認を確実にするために、あらゆるchrX領域を除去し、平均リード深さが30未満である領域を取り除いた。各バリアントおよびそのすぐ上流および下流のヌクレオチドは、64個のトリヌクレオチドコンテクストのうちの1つの中にある。中間のヌクレオチドを3つの他の塩基へと変異させる場合、全体で64×3=192個のトリヌクレオチド構成が可能である。トリヌクレオチド構成およびそれらの逆相補配列は等価であるので、実質的に96個のトリヌクレオチドコンテクストがある。トリヌクレオチドコンテクストは変異率に対して非常に強い影響があり、GCバイアス遺伝子変換に対する影響がより小さいことが観察され、これにより、これらの変数をモデル化するためにトリヌクレオチドコンテクストが有効になる。

【０４３６】

これらのイントロン領域内で、アレル頻度の4つのカテゴリ(シングルトン、シングルトンより多い～0.01%、0.01～0.1%、>0.1%)および192個のトリヌクレオチドコンテクストに基づいて、126136個のExAC/gnomADエクソンから各バリアントを取り込み、それらを4×192個のカテゴリへと分離した。そのトリヌクレオチドコンテクスト(イントロン配列の中の各ヌクレオチドを3つの異なる方式で置換することによって得られる)で潜在的なバリアントの総数を割ることによって、4×192個のカテゴリ(アレル頻度×トリヌクレオチドコンテクスト)の各々において観察されるバリアントの数を正規化した。192個のトリヌクレオチドコンテクストの各々に対して、タンパク質コーディング選択がないとき、4つのアレル頻度カテゴリの各々に該当するバリアントの予想される割合をこうして得た。これは、トリヌクレオチドコンテクストの差による、変異率、GCバイアス遺伝子変換、および遺伝的浮動の影響を暗黙的にモデル化する(補足テーブル7)。

【０４３７】

各アレル頻度カテゴリにおける予想されるミスセンス:同義比を得るために、一塩基置換によって入手可能なヒトゲノムにおける潜在的な同義変異およびミスセンス変異の総数をカウントし、それらの各々を192個のトリヌクレオチドコンテクストのうちの1つに割り当てた。各コンテクストに対して、4つのアレル頻度カテゴリの各々に該当することが予想されるバリアントの数を計算するために、4×192個のテーブルを使用した。最後に、192個のトリヌクレオチドコンテクストにわたる同義バリアントおよびミスセンスバリアントの数を合計して、4つのアレル頻度カテゴリの各々における同義バリアントとミスセンスバリアントの合計の予想される数を得た(図51および補足テーブル8(図59))。

【０４３８】

予想されるミスセンス:同義比が2.46:1であったシングルトンバリアントを除き、予想

されるミスセンス:同義比は、アレル頻度スペクトラムにわたってほぼ一定であり、自然選択がない場合にde novoバリアントについて予想される2.23:1という比に近かった。このことは、タンパク質コーディング選択圧力とは独立の要因(変異率、遺伝子変換、遺伝的浮動)の活動により、ExAC/gnomADにおけるシングルトンアレル頻度カテゴリのバリアントが、デフォルトでde novo変異より約10%高いミスセンス:同義比を有することが予想されることを示す。これを訂正するために、アレル頻度分析において、シングルトンに対するミスセンス:同義比を10%だけ下に調整した(図49A、図49B、図49C、図49D、および図49E、ならびに図50A、図50B、図50C、および図50D)。この小さい調整は、霊長類および他の哺乳類において存在する一般的なヒトバリアントに対する推定されるミスセンス枯渇率をおよそ3.8%低下させた(図49A、図49B、図49C、図49D、および図49E、ならびに図50A、図50B、図50C、および図50Dに示される)。シングルトンバリアントに対するより高いミスセンス:同義比は、トランスバージョン変異(これはミスセンス変化を作り出す可能性がより高い)より高い変異率が原因で、トランジション変異(これは同義変化を作り出す可能性がより高い)がより高いアレル頻度を有することが原因である。

【０４３９】

その上、このことは、2.23:1というde novo変異について予想される比を超える、ExAC/gnomADにおけるシングルトンバリアントに対する2.33:1という観察されるミスセンス:同義比を説明する。ミスセンス:同義比に対するアレル頻度スペクトラムの影響を考慮した後で、これは実際には予想と比較して5.3%のシングルトンバリアントの枯渇を反映しており、これはおそらく、de novo優性遺伝モードを持つ病原性ミスセンス変異に対する選択によるものである。実際に、機能喪失の確率が高い(pLI>0.9)ハプロ不全遺伝子だけを考慮するとき、ExAC/gnomADシングルトンバリアントに対するミスセンス:同義比は2.04:1であり、ハプロ不全遺伝子内で約17%前後の枯渇率を示す。この結果は、ある程度の不完全な浸透を仮定するとき、ミスセンス変異の20%が機能喪失変異と等価であるという以前の推定と合致する。

【０４４０】

変異率の大きな違いによる、ヒトアレル頻度スペクトラムにわたるCpGバリアントおよび非CpGバリアントに対するミスセンス:同義比も具体的に調査した(図52A、図52B、図52C、および図52D)。CpG変異と非CpG変異の両方に対して、一般的なチンパンジー多型と同一状態であるヒトバリアントは、アレル頻度スペクトラムにわたってほぼ一定のミスセンス:同義比を有することを確認した。

【０４４１】

［他の種における多型と同一状態であるヒトミスセンスバリアントの枯渇率］

他の種からのバリアントがヒトにおいて一般的なアレル頻度(>0.1%)で耐えられるかどうかを評価するために、他の種における変異と同一状態であったヒトバリアントを特定した。バリアントの各々に対して、ヒト集団におけるアレル頻度(シングルトン、シングルトンより多い～0.01%、0.01%～0.1%、>0.1%)に基づいて、それらを4つのカテゴリのうちの1つに割り当て、稀なバリアント(<0.1%)と一般的なバリアント(>0.1%)との間でのミスセンス:同義比(MSR)の低下を推定した。一般的なヒトアレル頻度(>0.1%)における同一状態のミスセンスバリアントの枯渇率は、ヒトにおいて一般的なアレル頻度では自然選択により除去されるのに十分有害な、他の種からのバリアントの割合を示す。

【０４４２】

【数４２】

$$\%\,\text{depletion} = \frac{\text{MSR}_{\text{rare}} - \text{MSR}_{\text{comm}}}{\text{MSR}_{\text{rare}}}$$

【０４４３】

ミスセンス:同義比および枯渇の百分率は、種毎に計算され、図50Bおよび補足テーブル

2に示されている。加えて、一般的なチンパンジーバリアント(図49A)、チンパンジーシングルトンバリアント(図49C)、および哺乳類バリアント(図50A)に対して、稀なバリアントと一般的なバリアントとの間のミスセンス:同義比の差が有意であったかどうかを検定するために、2×2の分割表上で相同性のカイ二乗(χ2)検定を実行した。

【０４４４】

シーケンシングは大型類人猿多様性プロジェクトからの限られた数の個体に対してのみ実行されたので、一般のチンパンジー集団において稀(<0.1%)または一般的(>0.1%)であったサンプリングされたバリアントの割合を推定するために、ExAC/gnomADからのヒトアレル頻度スペクトラムを使用した。ExAC/gnomADアレル頻度に基づく24体の個体のコホートをサンプリングし、このコホートにおいて一度観察された、または一度より多く観察されたミスセンスバリアントを特定した。一度より多く観察されたバリアントは、一般の集団において一般的(>0.1%)である確率が99.8%であったのに対し、コホートにおいて一度だけ観察されたバリアントは、一般の集団において一般的である確率が69%であった。図49Bおよび図49Cにおいて、チンパンジーシングルトンバリアントの一部が稀な有害な変異であることの結果として、ヒトにおいて高いアレル頻度でシングルトンチンパンジーバリアントの枯渇が観察されるが、一般的なチンパンジーバリアントについてはそれが観察されないということが示される。24体の個体のコホートにおいて観察されるチンパンジーバリアントの概ね半分は一度だけ観察され、概ね半分は一度より多く観察された。

【０４４５】

より遠縁の哺乳類におけるミスセンスバリアントに対する観察される枯渇率は、よく保存されている、したがってより正確にアラインメントされている遺伝子の混乱をもたらす影響によるものではなかったことを確認するために、ヒトと比較して11種の霊長類および50種の哺乳類の複数配列アラインメントにおける50%を超える平均ヌクレオチド相同性を持つ遺伝子のみに制約して、上記の分析を繰り返した(補足テーブル3参照)。これは、結果に実質的な影響を与えることなく、分析から約7%のヒトタンパク質コーディング遺伝子を除去した。

【０４４６】

［霊長類、哺乳類、および遠縁の脊椎動物の間で固定された置換］

バリアントデータについての問題、または家畜化によるアーティファクト(dbSNPから選択された種の大半は家畜化されているので)により、dbSNP変異を使用した我々の結果が影響を受けなかったことを確実にするために、種内多型の代わりに近縁の種のペアからの固定された置換を使用した分析も繰り返した。枝長で測定される進化系統距離(場所当たりのヌクレオチド置換の平均の数)とともに、UCSCゲノムブラウザから100種の脊椎動物の種の進化系統樹をダウンロードした(http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/hg19.100way.commonNames.nh)。さらなる分析のために、近縁の種のペア(枝長<0.25)を選択した。近縁の種のペア間の固定された置換を特定するために、ヒトとの99種の脊椎動物ゲノムの複数配列アラインメントのための、ならびにヒトとの19種の哺乳類(16種の霊長類)ゲノムのアラインメントのための、コーディング領域をUCSCゲノムブラウザからダウンロードした。追加の19種の哺乳類の複数種アライメントは、ボノボなどの霊長類の種の一部が99種の脊椎動物アラインメントにおいて存在しなかったので必要であった(http://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz20way/alignments/knownCanonical.exo nNuc.fa.gz)。全体で、図50Dおよび補足テーブル4に列挙されるように、5つの霊長類ペアを含む、近縁の種の15個のペアを得た。

【０４４７】

正規のコーディング領域内でのヒトとの19種の哺乳類ゲノムまたは99種の脊椎動物ゲノムの複数配列アライメントを取り込み、補足データファイル2において列挙される、脊椎動物の各々の選択されたペア間でのヌクレオチド置換を得た。コドンの中の他の2つのヌクレオチドがヒトと他の種との間で変わらなかったことを条件とし、かつ基準方向と代替方向のいずれかのバリアントを受け入れて、これらの置換がヒトゲノムにマッピングされた。関連する種のペアからの固定された置換と同一状態であったヒトバリアントを使用し

て、稀な(<0.1%)アレル頻度カテゴリのバリアントと一般的な(>0.1%)アレル頻度カテゴリのバリアントに対するミスセンス:同義比を計算し、補足テーブル4において示されるように、負の選択のもとで固定された置換の割合を得た。

【０４４８】

［ヒト、霊長類、哺乳類、および他の脊椎動物に対する多型データのClinVar分析］

他の種と同一状態であるバリアントの臨床上の影響を調査するために、ClinVarデータベース(2017年11月2日に発表されたftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/clinvar_20171029.vcf.gz)12のリリースバリアントサマリ(release variant summary)をダウンロードした。このデータベースは、hg19ゲノムビルド上の324698個のバリアントを含み、そのうち122884個がタンパク質コーディング遺伝子の我々のリストにマッピングするミスセンス一塩基バリアントであった(補足テーブル9)。ClinVarデータベースの中のバリアントの大半はミスセンスの結果をもたらさず、除外された。次に、矛盾する病原性の解釈を持つバリアントをフィルタリングし、良性、良性である可能性が高い、病原性、および病原性である可能性が高いアノテーションを伴うバリアントのみを残した。良性のアノテーションおよび良性である可能性が高いというアノテーションを持つバリアントを単一のカテゴリへと統合し、病原性のアノテーションまたは病原性である可能性が高いというアノテーションを持つバリアントも統合した。補足テーブル9に示されるフィルタリングステップの後で、全体で病原性カテゴリの中の24853個のバリアントおよび良性カテゴリの中の17775個のバリアントがあり、残りは有意性が知られていないまたは矛盾するアノテーションを伴うバリアントであるので、除外された。

【０４４９】

ヒト集団におけるClinVarミスセンスバリアントに対する基準を得るために、ExAC/gnomADアレル頻度からサンプリングされた30人の個人のコホートにおいてClinVarミスセンスバリアントを調査した。このコホートサイズは、霊長類多様性プロジェクト研究においてシーケンシングされた個人の数を概ね反映するように選ばれた。100個のそのようなシミュレーションからの、30人のヒトのコホートにおける病原性バリアントと良性バリアントの平均の数を報告する(図49E)。専門家は、ClinVarにおいて一般的なヒトバリアントを良性の結果で系統的にアノテートしてきたので、この精選のバイアスを避けるためにアレル頻度が1%より高いバリアントを除外した。

【０４５０】

霊長類、哺乳類、および他の脊椎動物における変異と同一状態であったClinVarバリアントを分析した。各種に対する良性バリアントおよび病原性バリアントの数が補足テーブル10に示されている。ヒト、霊長類、およびより遠縁の哺乳類において存在したClinVarバリアントの数の概要が、良性バリアントと病原性バリアントの比の差についての相同性のカイ二乗(χ2)検定からの結果とともに、図49Eおよび図50Bにおいて示されている。

【０４５１】

［モデル訓練のための良性バリアントの生成］

ヒト集団において一般的なバリアントは、創始者効果または平衡選択の稀な事例を除いて大部分が中立的であり、これにより、それらのバリアントは、人の解釈によるバイアスの影響を受けていない機械学習のための良性訓練データセットとして適切なものになる。フィルタを通過しなかったバリアントを除いて、ExAC/gnomADデータベース(リリースv2.0)からの123136個のエクソンからアレル頻度データを使用し、正規のタンパク質コーディング転写産物内で全体の集団アレル頻度が0.1%以上である83546個のミスセンスバリアントが残った。

【０４５２】

霊長類において存在するバリアントは大部分がヒトにおいて良性であることを示す我々の以前の結果に基づいて、一般的なヒトバリアント(>0.1%のアレル頻度)、大型類人猿多様性プロジェクトおよび追加の霊長類シーケンシングからのチンパンジー、ボノボ、ゴリラ、およびオランウータンからのバリアント、ならびに、dbSNPからのアカゲザル、オランウータン、およびマーモセットのバリアントを備える、機械学習のための良性訓練デー

タセットを作成した。一実装形態によれば、全体で、良性訓練セットに301690個の固有の霊長類バリアントが追加された。各源が寄与した良性訓練バリアントの数が補足テーブル5に示されている。

【０４５３】

注意すべき点は、大半の霊長類バリアントはそれらのそれぞれの集団において一般的であり、少数派が稀なバリアントであるということである。ヒト以外の霊長類の種は、シーケンシングされた個体の数が限られていたので、確認されたバリアントのセットは全般に、一般的な変異を表すことが予想される。実際に、霊長類の種の各々からのバリアントに対するミスセンス:同義比は、de novo変異に対する予想される2.23:1の比の半分未満であることが見出され、これらの大半が選択のふるいにすでにかけられてきた一般的なバリアントであることを示している。その上、チンパンジーのコホートに対して、確認されたバリアントの約84%が、それらのそれぞれの集団において一般的なアレル頻度(>0.1%)で存在することが推定された。新しく見つかったミスセンス変異の約50%が一般的なヒトアレル頻度(>0.1%)において純化選択によってフィルタリングされるので(図49A)、この数字は、観察された霊長類変異と同一状態であるヒトミスセンスバリアントの8.8%という観察された枯渇率を説明する、約16%の稀なバリアントと一致している(図49D)。

【０４５４】

ヒトミスセンス変異の約20%が機能喪失と等価であるという推定を適用すると、霊長類バリアントは3.2%の完全に病原性の変異と、91.2%の良性の変異(>0.1%のアレル頻度に耐える)と、完全には遺伝子の機能を失わず一般的なアレル頻度(>0.1%)で除去されるほど十分に有害ではない5.6%の中間的な変異とを含むことが予想される。この訓練データセットは不完全であることが知られているが、深層学習ネットワークの分類の正確さは、一般的なヒトバリアントと霊長類バリアントの両方を備える良性訓練データセット上で訓練されたとき、一般的なヒトバリアントのみの場合と比較してはるかに良好であった。したがって、現在の分類の正確さでは、利用可能な訓練データの量はより強い制限であるように見える。より多数の個体が各霊長類の種においてシーケンシングされるにつれて、より高い割合の一般的な霊長類バリアントを含む訓練データセットを準備することが可能になり、訓練データセットにおける病原性バリアントからの汚染が減り、分類性能がさらに改善する。

【０４５５】

［良性訓練データセットを補足するためのラベリングされていないバリアントの生成］

すべての潜在的なミスセンスバリアントが、正規のコーディング領域の各塩基場所から、その場所におけるヌクレオチドを他の3つのヌクレオチドで置換することによって生成された。ExAC/gnomADからの123136個のエクソンにおいて観察されたバリアントと、開始コドンまたは終止コドンにおけるバリアントを除外した。全体で、68,258,623個のラベリングされていないバリアントが生成された。ラベリングされていないバリアントの各々を、96個の異なるトリヌクレオチドコンテクストカテゴリのうちの1つに割り当てた。トリヌクレオチドコンテクストによって良性データセットの中のバリアントと一致する、このラベリングされていないデータセットからバリアントをサンプリングし、良性の訓練例とラベリングされていない訓練例を区別するように分類器を訓練することによって、半教師ありの手法を使用して深層学習ネットワークを訓練した。

【０４５６】

［ラベリングされていないバリアントの追加のフィルタリング］

良性バリアントおよびラベリングされていないバリアントの例をフランキングアミノ酸配列とともに提示することによって、深層学習ネットワークは、変異に対して高度に耐性のないタンパク質の領域を学習する。しかしながら、タンパク質配列の領域に一般的なバリアントがないことは、強い純化選択によるものであることがあり、または、バリアントが領域においてコールされるのを妨げる技術的なアーティファクトによるものであることがある。後者を訂正するために、ExAC/gnomADデータセットが1より小さい平均カバレッジを有していた領域から、良性データセットとラベリングされていないデータセットの両方

からのバリアントを除去した。同様に、ラベリングされていないバリアントを訓練の間に良性データセットの中の霊長類バリアントと照合するとき、霊長類が複数配列アラインメントにおいてヒトとのオーソロガスなアラインメント可能な配列を有しなかった領域から、ラベリングされていないバリアントを除外した。

【０４５７】

［妥当性確認および検定のための保留された霊長類バリアント、ならびに影響を受けている個人および影響を受けていない個人からのde novoバリアント］

深層学習ネットワークの妥当性確認および検定のために、妥当性確認および検定のために10000個の霊長類バリアントの2つのセットをランダムにサンプリングし、これらについては訓練を保留した。霊長類バリアントの残りは、一般的なヒトバリアント(>0.1%のアレル頻度)とともに、深層学習ネットワークを訓練するための良性データセットとして使用された。加えて、妥当性確認セットおよび検定セットのために、保留された霊長類バリアントと照合された10000個のラベリングされていないバリアントの2つのセットもサンプリングした。

【０４５８】

2つのセットの中のバリアントを区別するためのネットワークの能力を測ることによって、訓練の過程の間に深層学習ネットワークの性能を監視するために、妥当性確認セットの中の10000個の保留された霊長類バリアントを使用し、10000個のラベリングされていないバリアントを照合した。これにより、ネットワークの性能が飽和すると、訓練の停止点を決定して、過剰適応を避けることが可能になった。

【０４５９】

深層学習ネットワークならびに他の20個の分類器のベンチマークをとるために、検定データセットの中の10000個の保留された霊長類バリアントを使用した。異なる分類器は大きく変化するスコア分布を有していたので、これらのラベリングされていないバリアントを使用して、各分類器に対する50パーセンタイル閾値を特定した。方法間の公平な比較を確実にするために、その分類器に対して50パーセンタイル閾値で良性であるものと分類された、10000個の保留された霊長類バリアント検定セットの中のバリアントの割合について、各分類器のベンチマークをとった。

【０４６０】

神経発達障害を持つ影響を受けている個人におけるde novoバリアントと、健康な対照群におけるde novoバリアントとを使用して、臨床上の環境において深層学習ネットワークの性能を評価するために、Deciphering Developmental Disorders(DDD)研究からのde novoバリアントと、Simons Simplex Collection(SSC)自閉症研究における健康な兄弟の対照群からのde novoバリアントとをダウンロードした。DDD研究はde novoバリアントに対する信頼性レベルを提供し、バリアントコーリングエラーによる潜在的な偽陽性として、閾値が0.1未満であるDDDデータセットからのバリアントを除外した。全体で、DDDの影響を受けている個人からの3512個のミスセンスde novoバリアントおよび健康な対照群からの1208個のミスセンスde novoバリアントがあった。

【０４６１】

疾患遺伝子候補の顔ぶれの中で、有意性が不確かな良性バリアントおよび病原性バリアントを区別するという現実世界の臨床上のシナリオをより良くモデル化するために、タンパク質切断変異だけから計算された(補足テーブル18)、DDD研究において疾患と関連付けられた($p<0.05$)605個の遺伝子内のde novoバリアントのみに分析を限定した。遺伝子固有の変異率および考慮される染色体の数を仮定して、de novo変異の予想される数のヌル仮説のもとで統計的有意性を計算することによって、タンパク質切断de novo変異の遺伝子固有のエンリッチメントを評価した。名目のP値が0.05未満である605個の遺伝子を選択した。605個の遺伝子内の同義de novo変異およびミスセンスde novo変異の余剰(図22A)を、観察されるde novo変異vs予想されるde novo変異のカウントの比として、ならびに、観察されるde novo変異から予想されるde novo変異を引いた差分として計算した。これらの605個の遺伝子内で、DDDの影響を受けている個人から380個のde novoミスセンス変異を観察

した(図22A)。我々自身の分類器を含む分類器の各々に対して、小さい割合のバリアントは予測を有せず、それは全般に、それらのバリアントが分類器によって使用される同じ転写産物モデルにマッピングしなかったからである。したがって、我々の深層学習ネットワークでは、DDDの影響を受けている個人からの362個のde novoミスセンス変異および健康な対照群からの65個のde novoミスセンス変異を使用して、図22A、図22B、図22C、図22D、および図22Eにおいて下流分析を実行した。

【０４６２】

［シーケンシングされた霊長類集団の数の増大に伴うすべての潜在的なヒトミスセンス変異の飽和］

504種の現存する霊長類の種において存在する一般的なバリアントによる、すべての7000万個の潜在的なヒトミスセンス変異の予想される飽和を調査した。各霊長類の種に対して、ヒトにおいて観察される一般的なミスセンスバリアントの数(アレル頻度が0.1%より高い約83500個のミスセンスバリアント)を4回シミュレートした。それは、他の霊長類の種と比べてヒトの個体当たりのバリアントの数が概ね半分であるように見え、ヒトミスセンスバリアントの約50%が0.1%を超えるアレル頻度において純化選択により除去されているからである(図49A)。96個のトリヌクレオチドコンテクストにおける一般的なヒトミスセンスバリアントの観察される分布に基づいて、シミュレートされたバリアントを割り当てた。たとえば、一般的なヒトミスセンスバリアントの2%が、CCC>CTGのトリヌクレオチドコンテクストからのものであった場合、シミュレートされるバリアントの2%がランダムにサンプリングされたCCG>CTG変異であったことを要求した。これは、トリヌクレオチドコンテクストを使用して、変異率、遺伝的浮動、および遺伝子変換バイアスの影響を考慮する効果を有する。

【０４６３】

図23Dの曲線は、各霊長類の種におけるすべての一般的なバリアント(>0.1%のアレル頻度)を確認していることを仮定して、504種の霊長類の種のいずれかに存在する一般的なバリアントによる約7000万個の潜在的なヒトミスセンス変異の累積的な飽和を示す。図49Aから、ヒトミスセンス変異の概ね約50%が、ヒトと他の霊長類の両方において、一般的なアレル頻度(>0.1%)に達するのを妨げるのに十分有害であるので、図23Dの曲線は、無害なヒトミスセンス変異の割合が、霊長類の種の数が増えるにつれて一般的な霊長類変異により飽和することを表す。504種の霊長類の種を用いると、無害なヒトミスセンス変異の大半が飽和し、無害なCpG変異は変異率がより高いことが原因ではるかに少数の種で飽和する。

【０４６４】

調査されたヒトのコホートのサイズを大きくすることで発見された一般的なヒトミスセンスバリアント(>0.1%のアレル頻度)の割合をモデル化するために(図36)、gnomADアレル頻度に従って遺伝子型をサンプリングした。発見された一般的なgnomADミスセンスバリアントの割合は、10から100000までの各サンプルサイズに対する100回のシミュレーションにわたって平均された。

【０４６５】

［二次構造および溶媒接触性の予測］

病原性予測のための深層学習ネットワークは、二次構造および溶媒接触性予測ネットワークのための19個の畳み込み層と、二次構造および溶媒接触性ネットワークの結果を入力として取り込む主病原性予測ネットワークのための17個の畳み込み層とを含む、全体で36個の畳み込み層を含む。大半のヒトタンパク質の結晶構造は知られていないので、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ために2つのモデルを訓練した。両方のモデルが、図6に示される同じネットワークアーキテクチャおよび入力を使用した。二次構造および溶媒接触性ネットワークへの入力は、99種の他の脊椎動物とのヒトの複数配列アラインメントからの保存情報を符号化する、長さ51×20個のアミノ酸の位置特定的頻度行列である。

【０４６６】

二次構造ネットワークは、αヘリックス(H)、βシート(B)、およびコイル(C)という3状態の二次構造を予測するように訓練される。溶媒接触性ネットワークは、埋もれている(B)、中間(I)、および露出している(E)という3状態の溶媒接触性を予測するように訓練される。両方のネットワークが一次配列のみを入力として取り込み、Protein DataBankにおける既知の結晶構造からのラベルを使用して訓練された。モデルは各アミノ酸残基に対して1つの状態を予測する。

【０４６７】

［二次構造および溶媒接触性の予測のためのデータ準備］

モデルを訓練するために、Protein Databankからの関連しない結晶構造を使用した。25%を超える配列相動性を持つアミノ酸配列が除去された。全体で、6367個のタンパク質配列が訓練のために使用され、400個が妥当性確認のために使用され、500個が検定のために使用された(補足テーブル13)。アミノ酸配列および二次構造と溶媒接触性ラベルを含む、訓練のために使用されたデータは、RaptorXウェブサイト:http://raptorx.uchicago.edu/download/から入手可能である。

【０４６８】

大半の解かれている結晶構造はヒト以外のタンパク質のものであるので、二次構造および溶媒モデルを事前訓練するために、RaptorXスイート(PSI-BLASTに基づく)を使用して関連する配列を取得した。それは、ヒトベースの複数配列アラインメントが一般に入手可能ではなかったからである。RaptorXからCNFsearch1.66_releaseツールを使用してタンパク質に対する複数配列アラインメントを生成し、99個の最も近いアラインメントから各場所におけるアミノ酸をカウントして位置特定的頻度行列を形成した。たとえば、1u71A.fastaタンパク質に対する複数配列アラインメントを読み出すためのRaptorXを使用する具体的なコマンドは次の通りであった。

```
% ./buildFeature-i 1u71A.fasta-c 10-o ./TGT/1u71A.tgt
% ./CNFsearch-a 30-q 1u71A
```

【０４６９】

データセットの中の各アミノ酸の場所に対して、51個のフランキングアミノ酸に対応する位置特定的頻度行列からウィンドウを取り込み、これを使用して長さ51のアミノ酸配列の中心にあるアミノ酸に対する二次構造または溶媒接触性のいずれかのためのラベルを予測した。二次構造および相対的な溶媒接触性のためのラベルは、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。二次構造および溶媒接触性ネットワークを病原性予測ネットワークの一部として組み込むために、ヒトベースの99種の脊椎動物の複数配列アラインメントから位置特定的頻度行列を計算した。これらの2つの方法から生成された保存行列は一般に類似しているが、パラメータ重みの精密な調整を可能にするために、病原性予測のための訓練の間に二次構造および溶媒接触性モデルを通じた逆伝播を可能にした。

【０４７０】

［モデルアーキテクチャおよび訓練］

タンパク質の二次構造および相対的な溶媒接触性を予測するように、2つの別々の深層畳み込みニューラルネットワークモデルを訓練した。2つのモデルは、同一のアーキテクチャおよび入力データを有するが、予測状態については異なる。最高の性能に向けてモデルを最適化するために、詳細なハイパーパラメータ探索を行った。病原性予測のための我々の深層学習ネットワークと、二次構造および溶媒接触性を予測するための深層学習ネットワークの両方が、画像分類における成功により広く採用されている残差ブロックのアーキテクチャを採用した。残差ブロックは、より前の層からの情報が残差ブロックをスキップすることを可能にするスキップ接続が散在する、反復する畳み込みのユニットを備える。各残差ブロックにおいて、入力層がまずバッチ正規化され、正規化線形ユニット(ReLU)を使用する活性化層がそれに続く。活性化は次いで1D畳み込み層を通される。1D畳み込み層からのこの中間の出力は、再びバッチ正規化およびReLU活性化され、別の1D畳み込み層がそれに続く。第2の1D畳み込みの終わりに、その出力を元の入力と合計して残差ブロッ

クにし、このことが、元の入力情報が残差ブロックをバイパスすることを可能にすることによってスキップ接続として活動する。著者により深層残差学習ネットワークと名付けられるそのようなアーキテクチャでは、入力は元の状態で保存され、残差接続にはモデルからの非線形の活性化がない状態に保たれ、より深いネットワークの効果的な訓練が可能になる。詳細なアーキテクチャは、図6および補足テーブル11(図7Aおよび図7B)および図12(図8Aおよび図8B)において提供される。

【０４７１】

残差ブロックに続いて、ソフトマックス層が各アミノ酸に対する3状態の確率を計算し、それら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。モデルは、ADAM最適化器を使用して、タンパク質配列全体に対する累積カテゴリクロスエントロピー損失関数(accumulated categorical cross entropy loss function)を用いて訓練される。ネットワークが二次構造および溶媒接触性について事前訓練されると、病原性予測ネットワークのための入力としてネットワークの出力を直接取り込む代わりに、ソフトマックス層の前の層を取り込み、それによって、より多くの情報が病原性予測ネットワークを通る。

【０４７２】

3状態二次構造予測モデルについて達成される最高の検定の正確さは79.86%(補足テーブル14)であり、DeepCNF model30により予測される最新の正確さと同様である。3状態溶媒接触性予測モデルに対する最高の検定の正確さは60.31%(補足テーブル14)であり、同様の訓練データセット上でRaptorXによって予測される現在の最高の正確さと同様である。結晶構造を有していた約4000個のヒトタンパク質に対するDSSPでアノテートされた構造ラベルを使用するときと、予測された構造ラベルのみを使用する標準的なPrimateAIモデルを使用するときの、ニューラルネットワークの予測の比較も行った。DSSPでアノテートされたラベルを使用するとき、病原性予測の正確さにおけるさらなる改善は得られなかった(補足テーブル15)。

【０４７３】

［病原性予測のための深層学習モデルの入力特徴量］

病原性予測ネットワークのための訓練データセットは、フィルタリングの後で、385236個の良性とラベリングされたバリアントと、68258623個のラベリングされていないバリアントとを含む。各バリアントに対して、以下の入力特徴量を生成した。各バリアントの第1の入力特徴量は、バリアントの配列コンテクストを深層学習モデルに提供するための、長さ51のフランキングアミノ酸配列、すなわち、hg19の基準配列から得られたバリアントの各側への25個のアミノ酸である。全体で、このフランキング基準配列は長さが51個のアミノ酸である。経験的な観察結果を通じて、タンパク質配列のアミノ酸表現が、ヌクレオチドを使用してタンパク質コーディング配列を表現することより効果的であったことを発見した。

【０４７４】

第2の特徴量は、バリアントによって中心の場所において置換された代替アミノ酸を伴う、長さ51のフランキングヒトアミノ酸配列である。代替フランキング配列は、配列の中央の場所が基準アミノ酸の代わりに代替アミノ酸を含むことを除き、第1の特徴量における基準フランキング配列と同じである。基準ヒトアミノ酸配列と代替ヒトアミノ酸配列の両方が、長さ51×20のワンホット符号化されたベクトルへと変換され、各アミノ酸は、値0を伴う19個のアミノ酸および値1を伴う単一のアミノ酸のベクトルによって表される。

【０４７５】

11種の霊長類のための1つの位置特定的頻度行列(PFM)、霊長類を除く50種の哺乳類のための1つのPFM、および霊長類と哺乳類を除く38種の脊椎動物のための1つのPFMを含む、3つのPFMが、バリアントに対する99種の脊椎動物の複数配列アラインメントから生成される。各PFMはL×20の次元を有し、Lはバリアントの周りのフランキング配列の長さである(我々の事例では、Lは51個のアミノ酸を表す)。

【０４７６】

事前訓練された3状態二次構造および3状態溶媒接触性ネットワークへの入力のために、

やはり長さが51であり深さが20である、すべての99種の脊椎動物に対する複数配列アラインメントから生成される単一のPFM行列を使用した。Protein DataBankからの既知の結晶構造についてネットワークを事前訓練した後で、二次構造および溶媒モデルに対する最後の2つの層が除去され(グローバル最大プーリング層および出力層)、以前の層の出力の51×40の形状が、病原性予測ネットワークに対する入力として使用された。パラメータを精密に調整するために、ネットワークの構造層を通じた逆伝播を許容した。

【０４７７】

［半教師あり学習］

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リングされたインスタンスとラベリングされていないインスタンスの両方を使用するので、訓練に利用可能な少量のラベリングされたデータしかない完全教師あり（completely supervised）学習アルゴリズムよりも高い性能を達成する分類器を生み出すことができる。半教師あり学習の背後にある原理は、ラベリングされたインスタンスだけ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、ラベリングされていないデータ内の固有の知識を活用できるということであり、それにより半教師あり学習の潜在的な利益がもたらされる。少量のラベリングされたデータから教師あり分類器により学習されるモデルパラメータは、ラベリングされていないデータによって、より現実的な分布(これは検定データの分布によく似ている)に向かって導かれ得る。

【０４７８】

バイオインフォマティクスにおいて広まっている別の課題はデータ不均衡の問題である。データ不均衡現象は、予測されるべきクラスのうちの1つがデータにおいて過小評価されているときに生じ、それは、そのクラスに属するインスタンスが稀であり(注目に値する事例)、または取得が難しいからである。皮肉なことに、少数派のクラスが通常は最も学習に重要であり、それはそれらが特別な事例と関連付けられ得るからである。

【０４７９】

不均衡なデータ分布に対処するためのアルゴリズム手法は、分類器のアンサンブルに基づく。ラベリングされたデータの量が限られていることは、当然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どのような単一の構成分類器の性能も超える傾向がある。その上、アンサンブルは通常、複数のモデルを学習することと関連付けられる労力およびコストの妥当性を立証する要因によって、単一の分類器から取得される予測の正確さを改善する。直観的には、いくつかの分類器を集約することはより優れた過剰適合の制御につながり、それは、個々の分類器の高い変動制を平均化することは分類器の過剰適合も平均化するからである。

【０４８０】

高い信頼性でラベリングされた病原性バリアントの適切なサイズのデータセットがなかったので、半教師あり学習の戦略を追求した。ClinVarデータベースは300000個のエントリを有するが、有意性が不確かなバリアントを除去した後で、病原性の解釈が矛盾しないミスセンスバリアントは約42000個しか残らなかった。

【０４８１】

系統的な検討により、これらのエントリがアノテートされた病原性を支持するのに十分な臨床的エビデンスを持っていないことも発見された。その上、人により精選されたデータベースの中のバリアントの大半は、遺伝子の非常に小さい集合の中にある傾向があり、これにより、それらのバリアントが、一般的なヒトバリアントまたはチンパンジーとヒトとで固定された置換を使用してゲノムワイドで確認された良性訓練データセットの中のバリアントと一致しなくなる。データセットがどれだけ異なるように確認されたかを考慮すると、人により精選されたバリアントを病原性セットとして、およびゲノム全体の一般的なバリアントを良性セットとして用いて、教師あり学習モデル訓練することは、重大なバイアスをもたらす可能性が高い。

【０４８２】

バイアスを除去するように注意深く照合された、ラベリングされた良性バリアントのセ

ットとバリアントのラベリングされていないセットとを区別するように、深層学習ネットワークを訓練した。一実装形態によれば、385236個のラベリングされた良性バリアントのセットは、ExAC/gnomADデータベースからの一般的なヒトバリアント(>0.1%アレル頻度)およびヒト以外の霊長類の6つの種からのバリアントを含んでいた。

【０４８３】

(変異率、遺伝的浮動、および遺伝子変換を考慮するために)トリヌクレオチドコンテクストでの良性バリアントとの一致を条件とし、バリアント確認に対するアラインメント可能性およびシーケンスカバレッジの影響を調整して、ラベリングされていないバリアントのセットをサンプリングした。ラベリングされていないバリアントの数はラベリングされた良性バリアントを大きく超えるので、ラベリングされた良性バリアントの同じセットおよびラベリングされていないバリアントの8つのランダムにサンプリングされたセットを使用する8つのモデルを訓練し、それらの予測の平均をとることによって、コンセンサス予測を得た。

【０４８４】

半教師あり学習を選ぶ動機は、人により精選されたバリアントデータベースが、信頼できずノイズが多く、特に、信頼性のある病原性バリアントを欠いているということにある。gnomADからの一般的なヒトバリアントおよび霊長類バリアントから、信頼性のある良性バリアントのセットを得た。病原性バリアントについては、未知のバリアントのセット(臨床的な有意性がアノテートされていないVUSバリアント)からの病原性バリアントを最初にサンプリングすることに対して、反復的バランスサンプリング(iterative balanced sampling)手法を採用した。

【０４８５】

サンプリングバイアスを下げるために、良性訓練バリアントの同じセットおよび病原性バリアントの8つの異なるセットを使用する、8つのモデルのアンサンブルを訓練した。最初に、病原性バリアントを表現するために未知のバリアントをランダムにサンプリングした。次に、反復的に、モデルのアンサンブルが、前の訓練サイクルに関与していなかった未知のバリアントのセットをスコアリングするために使用される。次いで、上位のスコアリングされた病原性バリアントが、前のサイクルにおけるランダムな未知のバリアントの5%を置き換えるために取得される。必要とされるよりも25%多くの上位にスコアリングされる病原性バリアントを保持したので、8つのモデルに対するランダム性を高める未知のバリアントを置き換えるために、スコアリングされた病原性バリアントの8つの異なるセットをサンプリングできることに留意されたい。次いで、新しい病原性訓練セットが形成され、新しい訓練のサイクルが実行される。このプロセスは、最初のランダムにサンプリングされた未知のバリアントが、アンサンブルモデルによって予測される信頼性の高い病原性バリアントによりすべて置き換えられるまで繰り返される。図42は、反復的バランスサンプリングプロセスの例示である。

【０４８６】

［良性訓練セットおよび未知の訓練セットのバランスをとること］

良性バリアントと一致している未知のバリアントのサンプリング方式は、我々のモデル訓練のバイアスを低減するのに有用である。未知のバリアントがランダムにサンプリングされるとき、深層学習モデルはしばしば、偏った情報を抽出して自明解を提示する。たとえば、アミノ酸置換K→Mが良性バリアントより未知のバリアントにおいて頻繁に発生する場合、深層学習モデルはK→Mの置換を常に病原性として分類する傾向がある。したがって、2つの訓練セットの間でアミノ酸配列の分布のバランスをとることが重要である。

【０４８７】

CpGトランジションのようなより変異可能性の高いクラスは、一般的な良性バリアントにおいて大きな表現バイアスを有する。他の霊長類からのオーソロガスバリアントもヒトの変異率に従い、良性訓練セット全体における変異可能性の高いクラスのエンリッチメントを示唆する。未知のバリアントのサンプリング手順があまり制御されておらず、バランスがとれていない場合、深層学習モデルは、トランスバージョンまたは非CpGトランジシ

ョンなどのより出現しないクラスと比較して、CpGトランジションを良性として分類する可能性がより高い傾向がある。

【０４８８】

深層学習モデルが自明で非生物学的な解に収束するのを防ぐために、良性バリアントおよび未知のバリアントのトリヌクレオチドコンテクストのバランスをとることを考える。トリヌクレオチドは、バリアントの前の塩基、バリアントの基準塩基、およびバリアントの後の塩基によって形成される。そして、バリアントの基準塩基は他の3つのヌクレオチドへと変更され得る。全体で、64×3個のトリヌクレオチドコンテクストがある。

【０４８９】

［反復的バランスサンプリング］

サイクル1

各トリヌクレオチドコンテクストに対する良性バリアントの厳密な数と一致するように未知のバリアントをサンプリングした。言い換えると、最初のサイクルにおいて、バリアントのトリヌクレオチドコンテクストに関して良性訓練セットおよび病原性訓練セットを鏡写しにした。そのようなサンプリング方法の背後にある直観は、良性セットと未知のセットの間で変異率が同一であるバリアントの等しい表現があるということである。このことは、モデルが変異率に基づいて自明解に収束するのを防ぐ。

【０４９０】

サイクル2～サイクル20

サイクル2に対して、サイクル1からの訓練されたモデルを適用してサイクル1に関与していない未知のバリアントのセットをスコアリングし、上位の予測される病原性バリアントで未知のバリアントの5%を置き換えた。このセットは純粋にモデルによって生成され、このセットの中のトリヌクレオチドコンテクストに対するバランシングは適用しなかった。訓練に必要な未知のバリアントの残りの95%は、良性バリアントの中の各ヌクレオチドコンテクストのカウントの95%となるようにサンプリングされる。

【０４９１】

直観的には、サイクル1は完全に一致した訓練セットを使用するので、上位の予測される病原性バリアントはどのような変異率のバイアスも伴わずに生成される。したがって、このセットにおいてはどのようなバイアスも考慮する必要はない。データの残りの95%は依然として、モデルが自明解へ収束するのを防ぐために、トリヌクレオチドコンテクスト変異率のために制御される。

【０４９２】

各サイクルに対して、置き換えられた未知のバリアントの百分率は5%ずつ上昇する。サイクル3では、サイクル3のモデルからの上位の予測される病原性バリアントで未知のバリアントの5%を置き換えた。累積すると、病原性バリアントの割合は10%に上昇し、トリヌクレオチドコンテクストと鏡写しにされた未知のバリアントの割合は90%に下落する。サンプリングプロセスは残りのサイクルに対しても同様である。

【０４９３】

サイクル21

最後のサイクルであるサイクル21では、病原性訓練セット全体が、純粋に深層学習モデルから予測される上位の病原性バリアントからなる。各サイクルにおいて変異率のバイアスを明確に考慮してきたので、病原性バリアントは、訓練データとして使用するのに信頼性が高く、変異率のバイアスの影響を受けていない。したがって、訓練の最後のサイクルは、病原性予測のための最後の深層学習モデルを生み出す。

【０４９４】

［ラベリングされた良性訓練セットとラベリングされていない訓練セットの照合］

ラベリングされていないバリアントのバランスサンプリングが、バリアントの有害性に関連しないバイアスを除去するのに決定的に重要である。混乱をもたらす影響の適切な制御がないと、深層学習は容易に、不注意にもたらされたバイアスを選択してクラスを区別することがある。一般的なヒトバリアントは、CpGアイランド上のバリアントなどの、変

異可能性の高いクラスからのバリアントについてエンリッチされる傾向がある。同様に、霊長類多型はヒトの変異率にも従い、良性訓練セット全体における変異可能性の高いバリアントのエンリッチメントを示唆する。ラベリングされていないバリアントのサンプリング手順がよく制御されておらずバランスがとれていない場合、深層学習ネットワークは、バリアントを分類するために変異率のバイアスに頼る傾向があるので、トランスバージョンまたは非CpGトランジションなどのより出現しないクラスと比較して、CpGトランジションを良性として分類する可能性がより高い。我々は、96個のトリヌクレオチドコンテクスト(上で論じられた)の各々において、ラベリングされた良性バリアントと厳密に同じ数のラベリングされていないバリアントをサンプリングした。

【０４９５】

ラベリングされた良性データセットの中の霊長類バリアントに対してラベリングされていないバリアントを照合するとき、我々は、複数配列アラインメントにおいて霊長類の種がアラインメントされなかったヒトゲノムの領域から、ラベリングされていないバリアントが選択されるのを認めなかった。それは、その場所においてその霊長類の種の中のバリアントをコールすることが可能ではなかったからである。

【０４９６】

96個のトリヌクレオチドコンテクストの各々の中で、霊長類バリアントに対するシーケンシングカバレッジを訂正した。シーケンシングされるヒトの数が多いので、ヒト集団における一般的なバリアントは、シーケンシングカバレッジが低いエリアにおいても、それらが十分に確認されるほど十分に頻繁に観察される。同じことは霊長類バリアントに当てはまらず、それは、少数の個体しかシーケンシングされなかったからである。ExAC/gnomADエクソンにおけるシーケンシングカバレッジに基づいて、ゲノムを10個のビンへと分割した。各ビンに対して、ラベリングされた良性データセットvsラベリングされていないデータセットにおける霊長類バリアントの割合を測定した。シーケンシングカバレッジのみに基づいて、線形回帰を使用して、霊長類バリアントがラベリングされた良性データセットのメンバーである確率を計算した(図24)。ラベリングされた良性データセットにおける霊長類バリアントと照合すべきラベリングされていないバリアントを選択するとき、回帰係数を使用してその場所におけるシーケンシングカバレッジに基づいてバリアントをサンプリングする確率を重み付けた。

【０４９７】

［良性バリアントおよび未知のバリアントの生成］

ヒト集団における一般的なバリアント

最近の研究は、ヒト集団における一般的なバリアントが全般に良性であることを実証している。一実装形態によれば、gnomADは、正規のコーディング領域内でマイナーアレル頻度(MAF)が0.1%以上である90958個の非同義SNPを提供する。フィルタを通過したバリアントが保持される。インデルが除外される。開始コドンまたは終止コドンにおいて発生するバリアント、ならびにタンパク質切断バリアントが除去される。亜集団を精査すると、各亜集団内のMAFが0.1%以上であるミスセンスバリアントの総数は、一実装形態によれば245360個まで増える。これらのバリアントは、良性バリアントの訓練セットの一部を形成する。

【０４９８】

大型類人猿における一般的な多型

コーディング領域は高度に保存的であることが知られているので、多型が大型類人猿の集団において高い頻度で分離しているかどうかを仮定するのは簡単であり、多型は健康に対する軽度の影響も有し得る。大型類人猿ゲノムプロジェクトおよび他の研究からの、ボノボ、チンパンジー、ゴリラ、およびオランウータンの多型データは、dbSNPからのアカゲザルおよびマーモセットのSNPと統合された。

【０４９９】

未知のバリアントの生成

すべての潜在的なバリアントが、正規のコーディング領域の各塩基場所から、その場所

におけるヌクレオチドを他の3つのヌクレオチドに置換することによって生成される。新しいコドンが形成され、その場所におけるアミノ酸の潜在的な変更につながる。同義変化はフィルタリングされる。

【０５００】

gnomADデータセットにおいて観察されるバリアントが除去される。開始コドンまたは終止コドンにおいて発生するバリアント、ならびに終止コドンを形成するバリアントが除去される。複数の遺伝子アノテーションを持つSNPに対して、正規の遺伝子アノテーションが、SNPのアノテーションを表すために選択される。全体で、一実装形態によれば、68258623個の未知のバリアントが生成される。

【０５０１】

バリアントの追加のフィルタリング

ヒトゲノムの一部の領域では、リードをアラインメントするのが難しいことが知られている。それらの領域を含めると、訓練データセットおよび検定データセットに混乱をもたらす影響を引き起こす。たとえば、高い選択圧を受ける領域は、多型の数が限られる傾向がある。一方、シーケンシングが難しい領域もより少数の多型を有する。我々のモデルへのそのような混乱をもたらす入力を避けるために、gnomADによってシーケンシングされなかった遺伝子からのバリアントを除去した。

【０５０２】

一般に、良性バリアントは、複数の種にわたって保存される傾向があるよくシーケンシングされた領域において発見される。一方で、未知のバリアントは、いくつかのカバー率の低い領域を含むゲノム全体でランダムにサンプリングされる。これは、良性セットと未知のセットとの間に確認バイアスを引き起こす。バイアスを減らすために、gnomADにおいてリード深さが10未満であるバリアントを除去した。すべての哺乳類の種にわたるフランキング配列アラインメントにおいて10%を超える欠けているデータがあるすべてのバリアントも除去した。

【０５０３】

妥当性確認および検定のためのデータ

病原性モデルの妥当性確認および検定のために、一実装形態によれば、妥当性確認および検定のために、それぞれ10000個の良性バリアントの2つのセットを、良性バリアントの大きいプールからランダムにサンプリングした。良性バリアントの残りは、深層学習モデルを訓練するために使用される。これらのバリアントは特に、方法間の公平な比較を確実にするためにオーソロガスな霊長類バリアントからサンプリングされ、それは、一部の方法が一般的なヒトバリアントについて訓練されるからである。一実装形態によれば、妥当性確認および検定のために別々に、10000個の未知のバリアントの2つのセットをランダムにサンプリングした。192個のトリヌクレオチドコンテクストの各々の中の未知のバリアントの数が、妥当性確認セットおよび検定セットに対するそれぞれの良性バリアントの数と一致することを確実にする。

【０５０４】

自閉症または発育不全障害(DDD)を持つ影響を受けている子供と、影響を受けていない兄弟のde novoバリアントを使用して、臨床上の環境において複数の方法の性能を評価した。全体で、一実装形態によれば、DDDの症例群からの3821個のミスセンスde novoバリアントがあり、自閉症の症例群からの2736個のミスセンスde novoバリアントがある。一実装形態によれば、影響を受けていない兄弟について1231個のミスセンスde novoバリアントがある。

【０５０５】

［深層学習ネットワークアーキテクチャ］

病原性予測ネットワークは、二次構造および溶媒接触性ネットワークを介して、5つの直接入力および2つの間接入力を受け取る。5つの直接入力は、長さ51個のアミノ酸配列×深さ20(20個の異なるアミノ酸を符号化する)であり、バリアントを伴わない基準ヒトアミノ酸配列(1a)と、バリアントで置換された代替ヒトアミノ酸配列(1b)と、霊長類の種の複

数配列アラインメントからのPFM(1c)と、哺乳類の種の複数配列アラインメントからのPFM(1d)と、より遠縁の脊椎動物の種の複数配列アラインメントからのPFM(1e)とを備える。二次構造および溶媒接触性ネットワークは各々、複数配列アラインメント(1f)および(1g)からのPFMを入力として受け取り、主な病原性予測ネットワークへの入力として出力を提供する。二次構造および溶媒接触性ネットワークは、Protein DataBankのための既知のタンパク質結晶構造について事前訓練され、病原性モデル訓練の間の逆伝播を可能にする。

【０５０６】

5つの直接入力チャネルは、線形活性化を伴う40個のカーネルのアップサンプリング畳み込み層を通される。ヒト基準アミノ酸配列(1a)は、霊長類、哺乳類、および脊椎動物の複数配列アラインメントからのPFMと統合される(マージ1a)。同様に、ヒト代替アミノ酸配列(1b)は、霊長類、哺乳類、および脊椎動物の複数配列アラインメントからのPFMと統合される(マージ1b)。これは2つの並列な経路を作り出し、1つは基準配列のためのものであり、1つはバリアントで置換された代替配列のためのものである。

【０５０７】

基準チャネルと代替チャネルの両方の統合された特徴マップ(マージ1aおよびマージ1b)は、一連の6つの残差ブロックを通される(層2a～7a、マージ2aおよび層2b～7b、マージ2b)。残差ブロックの出力(マージ2aおよびマージ2b)は、サイズ(51,80)の特徴マップを形成するために一緒に連結され(マージ3a、マージ3b)、これは基準チャネルと代替チャネルからのデータを完全に混合する。次に、データは、セクション2.1において定義されるような2つの畳み込み層を各々含む一連の6つの残差ブロックを通じて(マージ3～9、層21、層34を除く層9～46)、または、1D畳み込みを通った後の1つおきの残差ブロックの出力を接続するスキップ接続を介して(層21、層37、層47)のいずれかで、ネットワークを並列に通るための2つの経路を有する。最終的に、統合された活性化(マージ10)が別の残差ブロックに供給される(層48～53、マージ11)。マージ11からの活性化は、フィルタサイズ1の1D畳み込みおよびシグモイド活性化に与えられ(層54)、次いで、ネットワークによるバリアント病原性の予測を表す単一の値を選ぶグローバル最大プーリング層に通される。モデルの概略的な図示が図3および補足テーブル16に示されている(図4A、図4B、および図4C)。

【０５０８】

［モデル概要］

バリアント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)モデルを開発した。モデルへの入力特徴量は、フランキングバリアントのタンパク質配列および保存プロファイルと、特定の遺伝子領域におけるミスセンスバリアントの枯渇率とを含む。深層学習モデルによって二次構造および溶媒接触性へバリアントによって引き起こされる変化を予測し、それを我々の病原性予測モデルへと統合した。モデルを訓練するために、ヒト亜集団の一般的なバリアントからの良性バリアントと、霊長類からのオーソロガスバリアントとを生成した。しかしながら、病原性バリアントに対する信頼性のある源が依然として欠けている。最初に、良性バリアントおよび未知のバリアントを用いてモデルを訓練し、次いで、半教師あり反復的バランスサンプリング(IBS)アルゴリズムを使用して、高い信頼性で予測される病原性バリアントのセットで未知のバリアントを徐々に置き換えた。最終的に、ヒトにおいて発育不全障害を引き起こすde novoバリアントを良性のバリアントから区別する際に、我々のモデルが既存の方法を上回ることを実証した。

【０５０９】

［残差ブロックの採用］

図17は残差ブロックを示す。病原性予測の我々の深層学習モデルと、二次構造および溶媒接触性を予測するための深層学習モデルの両方が、において最初に示された残差ブロックの定義を採用する。残差ブロックの構造は以下の図において示される。入力層は、まずバッチ正規化され、非線形活性化「ReLU」がそれに続く。活性化は次いで1D畳み込み層に通される。1D畳み込み層からのこの中間出力は、再びバッチ正規化およびReLU活性化され、別の1D畳み込み層が後に続く。第2の1D畳み込みの終わりにおいて、その出力を元の出

力と統合する。そのようなアーキテクチャでは、入力は元の状態に保たれ、残差接続はモデルの非線形活性化がない状態に保たれる。

【０５１０】

膨張/拡張畳み込みは、少数の訓練可能なパラメータで大きな受容野を可能にする。膨張/拡張畳み込みは、膨張畳み込み率または拡張係数とも呼ばれる、何らかのステップを用いて入力値をスキップすることによって、カーネルがその長さより大きいエリアにわたって適用されるような畳み込みである。膨張/拡張畳み込みは、畳み込み演算が実行されるときにより長い間隔の隣接する入力エントリ(たとえば、ヌクレオチド、アミノ酸)が考慮されるように、畳み込みフィルタ/カーネルの要素間に間隔を追加する。これにより、入力に長距離のコンテクスト依存性を組み込むことが可能になる。膨張畳み込みは、隣り合うヌクレオチドが処理される際に再使用するために部分的な畳み込み計算結果を保存する。

【０５１１】

［我々のモデルの新規性］

我々の方法は、3つの点でバリアントの病原性を予測するための既存の方法と異なる。第1に、我々の方法は、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のアーキテクチャを採用する。第2に、信頼性のある良性バリアントがgnomADからの一般的なヒトバリアントおよび霊長類バリアントから取得され、一方で、確実性の高い病原性訓練セットは、人により精選された同一のバリアントデータベースを使用したモデルの循環的な訓練および検定を避けるために、反復的バランスサンプリングおよび訓練を通じて生成される。第3に、二次構造および溶媒接触性のための深層学習モデルは、我々の病原性モデルのアーキテクチャへと統合される。構造および溶媒モデルから得られる情報は、特定のアミノ酸残基に対するラベル予測に限定されない。むしろ、リードアウト層が構造および溶媒モデルから除去され、事前訓練されたモデルが病原性モデルと統合される。病原性モデルを訓練する間、事前訓練された構造および溶媒層はまた、誤差を最小限にするために逆伝播する。これは、事前訓練された構造および溶媒モデルが、病原性予測問題に集中することを助ける。

【０５１２】

［二次構造および溶媒接触性モデルの訓練］

データ準備

タンパク質の3状態の二次構造および3状態の溶媒接触性を予測するために、深層畳み込みニューラルネットワークを訓練した。PDBからのタンパク質アノテーションが、モデルを訓練するために使用される。一実装形態によれば、配列プロファイルと25%を超える相同性を有する配列が除去される。一実装形態によれば、全体で、6293個のタンパク質配列が訓練のために使用され、392個が妥当性確認のために使用され、499個が検定のために使用される。

【０５１３】

タンパク質に対する位置特定的スコアリング行列(PSSM)保存プロファイルが、UniRef90を探すためにE値の閾値0.001および3回の反復という条件でPSI-BLASTを実行することによって生成される。あらゆる未知のアミノ酸ならびにその二次構造は、空白として設定される。また、すべてのヒト遺伝子に対する同様のパラメータ設定を用いてPSI-BLASTを実行し、それらのPSSM保存プロファイルを収集する。これらの行列は、構造モデルを病原性予測へと統合するために使用される。タンパク質配列のアミノ酸は次いで、ワンホット符号化ベクトルへと変換される。そして、タンパク質配列およびPSSM行列はL×20の行列へと形状変更され、Lはタンパク質の長さである。二次構造に対する3つの予測されるラベルは、ヘリックス(H)、βシート(B)、およびコイル(C)を含む。溶媒接触性に対する3つのラベルは、埋もれている(B)、中間(I)、および露出している(E)を含む。1つのラベルは1つのアミノ酸残基に対応する。ラベルは次元=3のワンホット符号化ベクトルとしてコーディングされる。

【０５１４】

モデルアーキテクチャおよび訓練

タンパク質の3状態の二次構造および3状態の溶媒接触性をそれぞれ予測するために、2つのエンドツーエンドの深層畳み込みニューラルネットワークモデルを訓練した。2つのモデルは同様の構成を有し、一方はタンパク質配列に対する、他方はタンパク質保存プロファイルに対する、2つの入力チャネルを含む。各入力チャネルは次元L×20を有し、Lはタンパク質の長さを示す。

【０５１５】

入力チャネルの各々は、40個のカーネルおよび線形活性化を伴う1D畳み込み層(層1aおよび層1b)を通される。この層は入力次元を20から40にアップサンプリングするために使用される。モデル全体ですべての他の層が40個のカーネルを使用することに留意されたい。2つの層(1aおよび1b)の活性化は、40個の次元の各々にわたって値を加算することによって一緒に統合される(すなわち、マージモード=「加算」)。マージノードの出力は、1D畳み込みの単一の層(層2)に、続いて線形活性化に通される。

【０５１６】

層2からの活性化は、上で定義されたような一連の9個の残差ブロック(層3～層11)を通される。層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、以下同様である。3つごとの残差ブロック(層5、層8、および層11)の出力を直接合計するスキップ接続もある。統合された活性化は次いで、ReLU活性化とともに2つの1D畳み込み(層12および層13)へと供給される。層13からの活性化はソフトマックスリードアウト層に与えられる。ソフトマックスは、所与の入力に対する3クラスの出力の確率を計算する。

【０５１７】

最良の二次構造モデルでは、1D畳み込みは1という膨張率を有する。溶媒接触性モデルに対して、最後の3つの残差ブロック(層9、層10、および層11)は、カーネルのカバレッジを上げるために2という膨張率を有する。タンパク質の二次構造は、近くにあるアミノ酸の相互作用に強く依存する。したがって、カーネルカバレッジがより高いモデルは、性能をわずかに改善する。一方で、溶媒接触性は、アミノ酸間の長距離の相互作用により影響を受ける。したがって、膨張畳み込みを使用する、カーネルのカバレッジが高いモデルに対して、その正確さはカバレッジが低いモデルの正確さより2%高い。

【０５１８】

以下の表は、一実装形態による、3状態二次構造予測モデルの各層に対する活性化およびパラメータについての詳細を提供する。

【０５１９】

【表２】

層	タイプ	カーネルの数、ウィンドウサイズ	形状	膨張率	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
マージ配列 +PSSM	マージ(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
マージ活性化	マージ-層 5、層 8 および層 11　モード=「加算」	-	(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

【０５２０】

一実装形態による、溶媒接触性モデルの詳細が以下の表に示されている。

【０５２１】

【表３】

層	タイプ	カーネルの数、ウィンドウサイズ	形状	膨張率	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
マージ配列+PSSM	マージ(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
マージ活性化	マージ-層 5、層 8 および層 11　モード=「加算」	-	(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

【０５２２】

特定のアミノ酸残基の二次構造クラスは、最大の予測されるソフトマックス確率によって決定される。モデルは、逆伝播を最適化するためのADAM最適化器を使用してタンパク質配列全体に対して累積カテゴリクロスエントロピー損失関数を用いて訓練される。

【０５２３】

3状態二次構造予測モデルに対する最高の検定の正確さは80.32%であり、同様の訓練データセット上でDeepCNFモデルによって予測される最新の正確さと同様である。

【０５２４】

3状態溶媒接触性予測モデルに対する最高の検定の正確さは64.83%であり、同様の訓練データセット上でRaptorXによって予測される現在の最高の正確さと同様である。

【０５２５】

以下で説明されるように、事前訓練された3状態二次構造および溶媒接触性予測モデルを、我々の病原性予測モデルへと統合した。

【０５２６】

［バリアントの病原性を予測するようにモデルを訓練すること］

病原性予測モデルの入力特徴量

上で論じられたように、病原性予測問題に対して、病原性モデルを訓練するための良性バリアント訓練セットおよび未知のバリアント訓練セットがある。各バリアントに対して、我々のモデルに供給するために以下の入力特徴量を準備した。

【０５２７】

各バリアントの第1の入力特徴量は、バリアントの配列コンテクストの深層学習モデルを提供するための、バリアントのフランキングアミノ酸配列、すなわち、hg19の基準配列から得られたバリアントの各側の25個のアミノ酸である。全体で、このフランキング基準配列は、51個のアミノ酸の長さを有する。

【０５２８】

第2の特徴量は、バリアントを引き起こした代替アミノ酸である。基準アミノ酸と代替アミノ酸のペアを直接提供する代わりに、代替フランキング配列をモデルに提供する。代替フランキング配列は、配列の中間の場所が基準アミノ酸ではなく代替アミノ酸を含むことを除き、第1の特徴量における基準フランキング配列と同じである。

【０５２９】

次いで、両方の配列が長さ51×20のワンホット符号化されたベクトルへと変換され、各アミノ酸は20個の0または1のベクトルによって表される。

【０５３０】

次いで、12種の霊長類のための1つの位置特定的重み行列(PWM)、霊長類を除く47種の哺乳類のための1つのPWM、および霊長類と哺乳類を除く40種の脊椎動物のための1つのPWMを含む、3つのPWMが、バリアントに対する99種の脊椎動物の複数配列アラインメント(MSA)から生成される。各PWMはL×20という次元を有し、Lはバリアントの周りのフランキング配列の長さである(我々の場合ではLは51個のアミノ酸を表す)。これは、種の各カテゴリにおいて見られるアミノ酸のカウントを備える。

【０５３１】

psi blastから、51個のアミノ酸のバリアントフランキング配列に対するPSSM行列も生成する。これは、病原性予測のための、3状態二次構造予測モデルおよび溶媒接触性予測モデルの統合に使用される。

【０５３２】

基準配列(入力1)、代替配列(入力2)、霊長類のためのPWM行列(入力3)、哺乳類(入力4)、脊椎動物(入力5)、ならびに3状態二次構造および溶媒接触性モデルからの情報を用いて、病原性モデルを訓練する。

【０５３３】

［深層学習モデルの訓練］

図19は、深層学習モデルワークフローの概要を提供するブロック図である。病原性訓練モデルは、5つの直接入力および4つの間接入力を備える。5つの直接入力特徴量は、基準配列(1a)、代替配列(1b)、霊長類保存率(1c)、哺乳類保存率(1d)、および脊椎動物保存率(1e)を含む。間接入力は、基準配列ベース二次構造(1f)、代替配列ベース二次構造(1g)、基準配列ベース溶媒接触性(1h)、および代替配列ベース溶媒接触性(1i)を含む。

【０５３４】

間接入力1fおよび1gに対して、ソフトマックス層を除く、二次構造予測モデルの事前訓練された層をロードする。入力1fに対して、事前訓練された層は、バリアントに対してPSI-BLASTによって生成されるPSSMとともにバリアントに対するヒト基準配列に基づく。同

様に、入力1gに対して、二次構造予測モデルの事前訓練された層は、PSSM行列とともに入力としてのヒト代替配列に基づく。入力1hおよび1iは、それぞれ、バリアントの基準配列および代替配列に対する溶媒接触性情報を含む同様の事前訓練されたチャネルに対応する。

【０５３５】

5つの直接入力チャネルは、線形活性化を伴う40個のカーネルのアップサンプリング畳み込み層を通される。層1a、層1c、層1d、および層1eは、40個の特徴量の次元にわたって加算された値と統合されて、層2aを作り出す。言い換えると、基準配列の特徴マップは、3つのタイプの保存率特徴マップと統合される。同様に、1b、1c、1d、および1eは、40個の特徴量の次元にわたって加算された値と統合されて、層2bを生成する。すなわち、代替配列の特徴量は、3つのタイプの保存率特徴量と統合される。

【０５３６】

層2aおよび2bは、ReLUの活性化を用いてバッチ正規化され、各々フィルタサイズ40の1D畳み込み層に通される(3aおよび3b)。層3aおよび層3bの出力は、互いに連結された特徴マップを伴う1f、1g、1h、および1iと統合される。言い換えると、保存プロファイルを伴う基準配列の特徴マップ、および保存プロファイルを伴う代替配列は、基準配列および代替配列の二次構造特徴マップ、ならびに基準配列および代替配列の溶媒接触性特徴マップと統合される(層4)。

【０５３７】

層4の出力は、6つの残差ブロック(層5、層6、層7、層8、層9、層10)を通される。最後の3つの残差ブロックは、カーネルにより高いカバレッジを与えるために、1D畳み込みに対して2という膨張率を有する。層10の出力は、フィルタサイズ1の1D畳み込みおよび活性化シグモイドを通される(層11)。層11の出力は、バリアントに対する単一の値を選ぶグローバル最大プールを通される。この値はバリアントの病原性を表す。病原性予測モデルの一実装形態の詳細が以下の表に示される。

【０５３８】

【表４Ａ】

層	タイプ	カーネルの数、ウィンドウサイズ	形状	膨張率	活性化
基準配列(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
霊長類保存率(1c)	畳み込み 1D	40,1	(51,40)	1	線形
哺乳類保存率(1d)	畳み込み 1D	40,1	(51,40)	1	線形
脊椎動物保存率(1e)	畳み込み 1D	40,1	(51, 40)	1	線形
基準配列ベース二次構造(1f)	入力層	-	(51,40)	-	-
代替配列ベース二次構造(1g)	入力層	-	(51,40)	-	-
基準配列ベース溶媒接触性(1h)	入力層	-	(51,40)	-	-
代替配列ベース溶媒接触性(1i)	入力層	-	(51,40)	-	-
基準配列マージ(2a)	マージ(モード=加算)(1a,1c,1d,1e)	-	(51,40)	-	-

【０５３９】

【表４Ｂ】

（【表４Ａ】の続き）

代替配列マージ(2b)	マージ(モード=加算)(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	マージ(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

【０５４０】

［アンサンブル］

一実装形態では、我々の方法の各サイクルに対して、同じ良性データセットおよび8つの異なる未知のデータセットで訓練する8つの異なるモデルを実行し、8つのモデルにわたって評価データセットの予測を平均した。未知のバリアントの複数のランダムにサンプリングされたセットがモデルに提示されると、サンプリングバイアスを減らしてよく制御することができる。

【０５４１】

さらに、アンサンブルのアンサンブルという手法の採用が、我々の評価データセットに対する我々のモデルの性能を改善することができる。CADDは、10個のモデルのアンサンブ

ルを使用して、バリアントをスコアリングするために10個すべてのモデルにわたる平均スコアを得る。ここで、同様のアンサンブル手法を使用することを試みた。1つのアンサンブルを使用して結果のベンチマークをとり、次いで、アンサンブルの数を増やして性能の向上を評価した。各アンサンブルは、同じ良性データセットおよび8つの異なる未知のデータセットで訓練する8つのモデルを有することに留意されたい。異なるアンサンブルに対して、乱数生成器のシード値は別個であるので、ランダムなバリアントのセットが互いに異なるように導かれる。

【０５４２】

一実装形態による詳細な結果が以下の表に示される。

【０５４３】

【表５】

アンサンブルの数	DDD データセットについての-log(pvalue)(アンサンブルの平均の平均)	DDD データセットについての-log(pvalue)(アンサンブルの平均の中央値)
1	3.4e-27	3.4e-27
5	2.74e-29	3.8e-29
10	2.98e-29	2.55e-29
15	4.06e-29	3.88e-29
20	3.116e-29	3.05e-29
25	3.77e-29	3.31e-29
30	3.81e-29	3.34e-29

【０５４４】

1つのアンサンブルと比較して、5つのアンサンブルおよび10個のアンサンブルが、DDDデータセットを使用して評価されるときにより大きなp値を生み出した。しかし、アンサンブルの数を増やすことでさらに性能が向上することはなく、アンサンブルに対する飽和を示している。アンサンブルは、多様な未知のバリアントを用いてサンプリングバイアスを減らす。しかしながら、良性クラスと病原性クラスとで192個のトリヌクレオチドコンテクストを照合することも必要とされ、これは我々のサンプリング空間をかなり制限し、早い飽和につながる。アンサンブルのアンサンブルという手法は、モデル性能を大きく改善し、モデルについての我々の理解をさらに深めると結論付ける。

【０５４５】

［病原性モデルを訓練するための早期打ち切り］

信頼性のあるアノテートされた病原性バリアントサンプルが欠けているので、モデル訓練のための打ち切り基準を定義するのは困難である。モデル評価における病原性バリアン

トの使用を避けるために、一実装形態では、オーソロガスな霊長類からの10000個の良性妥当性確認バリアントと、10000個のトリヌクレオチドコンテクストが照合された未知のバリアントとを使用した。モデルの各エポックを訓練した後、良性妥当性確認バリアントおよび未知の妥当性確認バリアントを評価した。妥当性確認バリアントセットの両方の確率分布の差を評価するために、ウィルコクソン順位和検定を使用した。

【０５４６】

検定のp値は、モデルが良性バリアントを未知のバリアントのセットから区別する能力の向上に伴って、より大きくなる。モデル訓練の任意の5つの連続するエポックの間に、2つの分布を区別するモデルの能力に改善が観察されない場合、訓練を打ち切る。

【０５４７】

事前に、10000個の保留された霊長類バリアントの2つの別個のセットを訓練から除外し、我々はこれらのセットを妥当性確認セットおよび検定セットと名付けた。モデル訓練の間の早期打ち切りを評価するために、10000個の保留された霊長類バリアントおよび10000個のトリヌクレオチドコンテクストについて照合されたラベリングされていないバリアントの妥当性確認セットを使用した。各訓練エポックの後で、ラベリングされた良性妥当性確認セットおよびラベリングされていない一致した対照群におけるバリアントを区別するための、深層ニューラルネットワークの能力を評価し、ウィルコクソン順位和検定を使用して予測されるスコアの分布の差を測った。5つの連続的な訓練エポックの後でさらなる改善が観察されないと、過剰適合を防ぐために訓練を打ち切った。

【０５４８】

［分類器の性能のベンチマーキング］

1つは一般的なヒトバリアントのみを用いて訓練され、1つは一般的なヒトバリアントと霊長類バリアントの両方を含む良性とラベリングされた完全なデータセットを用いて訓練された、2つのバージョンの深層学習ネットワークの分類の正確さを、以下の分類器、すなわちSIFT、PolyPhen-2、CADD、REVEL、M-CAP、LRT、MutationTaster、MutationAssessor、FATHMM、PROVEAN、VEST3、MetaSVM、MetaLR、MutPred、DANN、FATHMM-MKL_coding、Eigen、GenoCanyon、およびGERP++13,32-48に加えて評価した。他の分類器の各々のスコアを得るために、dbNSFP 49(https://sites.google.com/site/jpopgen/dbNSFP)からすべてのミスセンスバリアントに対するスコアをダウンロードし、10000個の保留された霊長類バリアント検定セット、およびDDD症例群vs対照群におけるde novoバリアントについて方法のベンチマークをとった。本明細書に含めるものには、SIFT、PolyPhen-2、およびCADD、ならびにREVELを選択した。それは、SIFT、PolyPhen-2、およびCADDについては、それらが最も広く使用されている方法であるからであり、REVELについては、様々な評価モードにわたって、評価した20個の既存の分類器の中で最良のものの1つとして傑出していたからである。評価したすべての分類器の性能が図28Aにおいて提供される。

【０５４９】

深層学習ネットワークの性能に対する、利用可能な訓練データサイズの影響を評価するために、385236個の霊長類バリアントと一般的なヒトバリアントのラベリングされた良性訓練セットからランダムにサンプリングすることによって、図6の各データ点において深層学習ネットワークを訓練した。分類器の性能におけるランダムなノイズを減らすために、初期パラメータ重みのランダムなインスタンス化を各回において使用してこの訓練手順を5回実行し、10000個の保留された霊長類バリアントとDDD症例群vs対照群データセットの両方についての中央値の性能を図6に示した。偶然にも、385236個のラベリングされた良性バリアントの完全なデータセットを用いた分類器の中央値の性能は、DDDデータセットについて本明細書の残りで使用したものよりわずかに高かった(ウィルコクソン順位和検定で$P<10^{-28}$ではなく$P<10^{-29}$)。各々の個別の霊長類の種からのバリアントが分類の正確さに寄与しており、一方で各々の個別の哺乳類の種からのバリアントが分類の正確さを下げることを示すために、一実装形態によれば、83546個のヒトバリアントと、各種に対する一定数のランダムに選択されたバリアントとを備える訓練データセットを使用して深層学習ネットワークを訓練した。一実装形態によれば、訓練セットに追加したバリアント

の一定の数(23380)は、ミスセンスバリアントの数が最小の種、すなわちボノボにおいて利用可能なバリアントの総数である。ノイズを減らすために、訓練手順を再び5回繰り返し、分類器の中央値の性能を報告した。

【０５５０】

［モード評価］

一実装形態では、反復的バランスサンプリング手順に続いて、深層学習モデルを21回のサイクルにわたり訓練した。我々の分類器の性能を評価するために、2つのタイプの評価を実行した。2つの尺度で我々のモデルとPolyphen2、SIFT、およびCADDの比較も行い、臨床的なアノテーションに対する我々のモデルの適用の可能性を評価した。

【０５５１】

方法1:良性検定セットの正確さ

一実装形態では、10000個の良性バリアントおよび未知のバリアントを、8つの異なる訓練されたモデルのアンサンブルを使用してそれらの予測される確率を計算することによって、評価した。上で言及された他の既存の方法によってスコアリングされる、それらの予測される確率も取得した。

【０５５２】

次いで、評価において使用される方法の各々に対する未知の検定バリアントにわたる予測される確率の中央値を取得した。中央値スコアを使用して、方法の各々によって使用される良性バリアントおよび病原性バリアントのアノテーションに応じ、スコアが中央値の上または下である良性バリアントの数を発見した。SIFT、CADD、および我々の方法は、病原性バリアントを1として、良性バリアントを0としてラベリングする。したがって、スコアが中央値の下である良性バリアントの数をカウントした。Polyphenは反対のアノテーションを使用し、中央値を上回る良性バリアントの数をカウントした。スコアが中央値の上/下である良性バリアントの数を良性バリアントの総数で割った比は、良性バリアントの予測の正確さを表す。

良性の正確さ=中央値を上回る(下回る*)良性バリアントの総数÷良性バリアントの総数

【０５５３】

この評価方法の背後にある我々の理論は、gnomADにおけるバリアントの選択圧力の分析に依拠している。gnomADにおけるシングルトンに対して、ミスセンスバリアントと同義バリアントの比は約2.26:1である。一方、gnomADにおける一般的なバリアント(MAF>0.1%)では、ミスセンス対同義比は約1.06:1である。これは、ランダムな未知のバリアントのセットから、約50%が自然選択によって排除されることが予想され、残りの50%は軽度であり集団において一般的になる傾向があることを示している。

【０５５４】

【表６】

方法	保留された良性セットの正確さ
Polyphen	0.74
SIFT	0.69
CADD	0.77
我々のモデル	0.85

【０５５５】
上の表に示されるように、我々の方法の性能は2番目に良い方法であるCADDより8%を超えて優れている。これは、良性バリアントを分類する我々のモデルの能力の大きな向上を示している。そのような実証は我々のモデルの能力を証明するが、以下の方法2は、臨床的な解釈のための臨床的なデータセットに対する我々のモデルの有用性を示す。
【０５５６】
方法2:臨床的なデータセットの評価
一実装形態では、発育不全障害(DDD)症例群-対照群データセットを含む、臨床的なデータセットに対してこれらの病原性予測方法を評価した。DDDデータセットは、影響を受けている子供からの3821個のde novoミスセンスバリアントおよび影響を受けていない兄弟からの1231個のde novoミスセンスバリアントを備える。我々の仮説は、影響を受けている子供からのde novoバリアントが影響を受けていない兄弟からのde novoバリアントより有害である傾向があるというものである。
【０５５７】
臨床的な検定データセットは病原性バリアントを明確にラベリングしないので、それらの方法の性能を測るために、(影響を受けている群および影響を受けていない群からの)de novoバリアントの2つのセットの分離を使用した。de novoバリアントのこれらの2つのセットの分離がどれだけ優れているかを評価するために、ウィルコクソン順位和検定を適用した。
【０５５８】

【表７】

方法	DDDデータセットについての-log10(p値)
Polyphen	15.02
SIFT	13.52
CADD	13.83
DL	28.35

【０５５９】
上の表によれば、我々の半教師あり深層学習モデルは、de novoバリアントの影響を受けているセットと影響を受けていないセットとを区別する際の性能がはるかに高い。これは、我々のモデルが臨床的な解釈に対して既存の方法より適切であることを示している。これはまた、ゲノム配列および保存プロファイルから特徴量を抽出するという全般的な手法が、人により精選されたデータセットに基づく人が加工した特徴量より優れていることを確証する。
【０５６０】
［10000個の霊長類バリアントの保留された検定セットについての良性予測の正確さ］
深層学習ネットワークならびに他の20個の分類器のベンチマークをとるために、検定データセットの中の10000個の保留された霊長類バリアントを使用した。異なる分類器は大きく変動するスコア分布を有していたので、各分類器に対する50パーセンタイル閾値を特定するために、トリヌクレオチドコンテクストにより検定セットと照合された、10000個のランダムに選択されたラベリングされていないバリアントを使用した。方法間の公平な比較を確実にするために、その分類器に対して50パーセンタイルの閾値で良性であると分類された、10000個の保留された霊長類バリアント検定セットの中のバリアントの割合について、各分類器のベンチマークをとった。
【０５６１】
良性バリアントを特定するために50パーセンタイルを使用することの背後にある我々の理論は、ExAC/gnomADデータセットの中のミスセンスバリアントに対して観察される選択圧力に基づく。シングルトンアレル頻度で発生するバリアントでは、ミスセンス:同義比は～2.2:1であるが、一般的なバリアント(>0.1%アレル頻度)では、ミスセンス:同義比は約1.06:1である。これは、ミスセンスバリアントの約50%が一般的なアレル頻度では自然選択により排除されることが予想され、残りの50%が遺伝子的浮動を介して集団において一般的になる可能性を有するのに十分軽度であることを示している。
【０５６２】
分類器の各々に対して、50パーセンタイル閾値を使用して良性であるものとして予測される保留された霊長類検定バリアントの割合が示されている(図28Aおよび補足テーブル17(図34))。
【０５６３】
［DDD研究からのde novoバリアントの分析］
DDDの影響を受けている個人におけるde novoミスセンスバリアントと、影響を受けてい

ない兄弟の対照群におけるde novoミスセンスバリアントとを区別する能力について、分類方法のベンチマークをとった。各分類器に対して、2つの分布に対する予測スコア間の差のウィルコクソン順位和検定からのp値を報告した(図28Bおよび図28Cおよび補足テーブル17(図34))。

【０５６４】

モデルの性能を分析するための我々の2つの尺度が異なる源および方法から導かれると仮定し、2つの異なる尺度についての分類器の性能が相関していたかどうかを検定した。実際に、我々はこれらの2つの尺度が相関していたことを発見し、保留された霊長類検定セットに対する良性分類の正確さと、DDD症例群vs対照群におけるde novoミスセンスバリアントに対するウィルコクソン順位和検定のp値との間で、spearman ρ=0.57(P<0.01)であった。これは、分類器のベンチマークをとることについて、保留された霊長類検定セットの正確さと、DDD症例群vs対照群のp値との間に、良い一致があることを示す(図30A)。

【０５６５】

さらに、深層学習ネットワークが疾患と関連付けられる遺伝子の発見を助け得るかどうかを検定した。観察されたde novo変異の数をヌル変異モデルのもとで予想される数と比較することによって、遺伝子におけるde novo変異のエンリッチメントを検定した。

【０５６６】

すべてのミスセンスde novo変異からの結果と、スコアが0.803より大きいミスセンス変異からの結果とを比較して、深層学習ネットワークの性能を調査した。すべてのミスセンスde novoを検定することにはデフォルトのミスセンス率を使用したが、フィルタリングされたミスセンスde novoを検定することにはスコアが0.803より大きいサイトから計算されたミスセンス変異率を使用した。各遺伝子には4つの検定が必要であり、1つはタンパク質切断エンリッチメントを検定し、1つはタンパク質を変化させるde novo変異のエンリッチメントを検定し、両方がDDDコホートだけに対して、および神経発達トリオシーケンシングコホートのより大きなメタ分析に対して検定される。タンパク質を変化させるde novo変異のエンリッチメントは、コーディング配列内のミスセンスde novo変異のクラスタリングの検定とともに、Fisherの方法によって合成された(補足テーブル20および21)。各遺伝子に対するp値は4回の検定の最小値から取られ、ゲノムワイド有意性が$P<6.757\times10^{-7}$と決定された(α=0.05、18500個の遺伝子、4回の検定)。

【０５６７】

［605個のDDD疾患関連遺伝子内での受信者操作特性曲線および分類の正確さの計算］

深層学習ネットワークが本当に同じ遺伝子内の病原性バリアントと良性バリアントとを区別していたかどうかを検定するために、de novo優性遺伝モードを伴う遺伝子における病原性を優先するのではなく、DDDコホートにおいてp値が0.05未満である(de novoタンパク質切断変異のみを使用して計算される)神経発達疾患と関連付けられた605の遺伝子のセットを特定した(補足テーブル18)。DDDデータセットおよび対照群データセットにおいて605個の遺伝子の中のバリアントの確率分布を分類器が分離する能力について、すべての分類器に対するウィルコクソン順位和のp値を報告する(図28Cおよび補足テーブル19(図35))。

【０５６８】

605個の遺伝子のこのセット内で、変異率だけから予想されるものの3倍のde novoミスセンスバリアントのエンリッチメント比を観察した。これは、影響を受けているDDD患者におけるde novoミスセンスバリアントが、約67%の病原性バリアントと33%のバックグラウンドバリアントを備え、一方で、健康な対照群におけるde novoミスセンスバリアントが、不完全な浸透の事例を除いて大半はバックグラウンドバリアントからなることを示している。

【０５６９】

病原性バリアントと良性バリアントを完璧に区別する分類器に対する最大の可能なAUCを計算するために、605個の遺伝子内の影響を受けている個人におけるde novoミスセンスバリアントの67%だけが病原性であり、残りがバックグラウンドであったことを考慮した

。受信者操作特性曲線を構築するために、病原性であるものとしてのde novo DDDバリアントの分類を真陽性のコールとして扱い、病原性であるものとしての健康な対照群におけるde novoバリアントの分類を偽陽性のコールとして扱った。したがって、完璧な分類器は、DDD患者におけるde novoバリアントの67%を真陽性として、DDD患者におけるde novoバリアントの33%を偽陽性として、対照群におけるde novoバリアントの100%を真陰性として分類する。受信者操作特性曲線の視覚化は、プロットの(0%,0%)の角および(100%,100%)の角へ直線によって接続された、真陽性率が67%であり偽陽性率が0%である単一の点を示すだけであり、良性変異と病原性変異の完璧な区別についての分類器に対する0.837という最大AUCが得られる(図30Bおよび補足テーブル19(図35))。

【０５７０】

合成されたDDDのデータセットおよび健康な対照群のデータセットにおける605個の遺伝子内での病原性バリアントの予想される割合を推定することによって、病原性バリアントと良性バリアントとをバイナリ閾値で分離するための深層学習ネットワークの分類の正確さを計算した。DDDデータセットは、予想を超える249個のde novoミスセンスバリアントの余剰を伴う379個のde novoバリアントを含み、対照群データセットは65個のde novoバリアントを含んでいたので、全体で444個のバリアントのうち249個の病原性バリアントを予測した(図22A)。この予想される比率に従って、444個のde novoミスセンスバリアントを良性または病原性カテゴリへと分離した各分類器に対する閾値を選択し、これをバイナリカットオフとして使用して各分類器の正確さを評価した。我々の深層学習モデルに対して、この閾値は、真陽性率が65%、偽陽性率が14%である0.803以上のカットオフにおいて達成された。DDDの個人における約33%のバックグラウンドバリアントの存在について調整された分類の正確さを計算するために、バックグラウンドであったde novo DDDバリアントの33%が、健康な対照群において観察したのと同じ偽陽性率で分類されるであろうと仮定した。これは、DDDデータセットにおける真陽性分類イベントの14%×0.33=4.6%が実際にはバックグラウンドバリアントからの偽陽性であることに対応する。深層学習ネットワークに対する調整された真陽性率を、(65%-4.6%)/67%=90%と推定する。深層学習ネットワークに対しては88%である、真陽性率および真陰性率の平均を報告する(図30Cおよび補足テーブル19(図35))。この推定は、神経発達障害において不完全な浸透が広く見られることにより、分類器の真の正確さを過小評価する可能性が高い。

【０５７１】

［ClinVar分類の正確さ］

既存の分類器の大半はClinVar上で訓練される。ClinVar上で直接訓練しない分類器も、ClinVar上で訓練される分類器からの予測スコアを使用することによって影響を受けることがある。加えて、一般的なヒトバリアントは良性のClinVarの結果に対して高度にエンリッチされ、それは、アレル頻度が、良性の結果をバリアントに割り当てるための基準の一部であるからである。

【０５７２】

我々は、ClinVarデータセットを分析に適したものにするために、2017年に追加されたClinVarバリアントだけを使用することによって、ClinVarデータセットにおける循環性を最小限にすることを試みた。それは、他の分類方法がその前の年に公開されたからである。2017年のClinVarバリアントの中でも、ExACにおいて一般的なアレル頻度(>0.1%)で存在するあらゆるバリアント、またはHGMD、LSDB、もしくはUniprotにおいて存在するあらゆるバリアントを除外した。すべてのそのようなバリアントをフィルタリングした後で、および有意性が不確かであり矛盾するアノテーションを伴うバリアントを除外した後で、ClinVarにおいて良性アノテーションを伴う177個のバリアントおよび病原性アノテーションを伴う969個のバリアントが残った。

【０５７３】

深層学習ネットワークと既存の方法の両方を使用してすべてのClinVarバリアントをスコアリングした。このデータセット内の良性バリアントおよび病原性バリアントの観察される比率に従って、ClinVarバリアントを良性カテゴリまたは病原性カテゴリに分離した

各分類器に対する閾値を選択し、これをバイナリカットオフとして使用して各分類器の正確さを評価した。各分類器に対する真陽性率および真陰性率の平均を報告する(図31Aおよび図31B)。ClinVarデータセットについての分類器の性能は、10000個の保留された霊長類バリアントに対する分類の正確さについての分類器の性能、またはDDD症例群vs対照群データセットに対するウィルコクソン順位和のp値についての分類器の性能のいずれとも大きく相関しなかった(図31Aおよび図31B)。

【０５７４】

我々は、既存の分類器は専門家の行動を正確にモデル化しているが、人の経験則は経験的なデータにおいて病原性変異と良性変異とを区別するのに完全に最適ではないことがあるという仮説を立てている。1つのそのような例はGranthamスコアであり、これはアミノ酸置換の相同性または非相同性を特徴付けるための距離の尺度を与える。完全なClinVarデータセット(約42000個のバリアント)内の病原性バリアントおよび良性バリアントに対する平均Granthamスコアを計算し、これを605個の遺伝子内の影響を受けているDDDの個人および影響を受けていない個人におけるde novoバリアントに対する平均Granthamスコアと比較した。影響を受けているDDDの個人における約33%のバックグラウンドバリアントの存在を訂正するために、DDD症例群vs対照群の間のGranthamスコアの差を50%増大させたが、それでもこれは、ClinVarにおける病原性バリアントと良性バリアントとの差より小さかった。1つの可能性は、専門家が、アミノ酸置換距離などの、測定しやすい尺度を重視しすぎている一方で、専門家にとって定量化がより難しいタンパク質構造などの要因を軽視しているということである。

【０５７５】

［深層学習モデルの解釈］

機械学習アルゴリズムが問題を解く手段を理解するのは難しいことが多い。バリアントの病原性を予測するために深層学習ネットワークが学習して抽出した特徴量を理解するために、深層学習ネットワークの初期層を視覚化した。事前訓練された3状態二次構造予測モデルの最初の3つの層(2つのアップサンプリング層とそれに続く第1の畳み込み層)内での異なるアミノ酸に対する相関係数を計算し、BLOSUM62行列またはGrantham距離と非常に似た特徴量を畳み込み層の重みが学習することを示した。

【０５７６】

異なるアミノ酸の間の相関係数を計算するために、二次構造モデルにおいて3つのアップサンプリング層(層1a、層1b、および層1c)が前にある第1の畳み込み層の重みから始めた。3つの層の間の行列乗算を実行し、次元が(20,5,40)である行列を得た。ここで、20はアミノ酸の数であり、5は畳み込み層のウィンドウサイズであり、40はカーネルの数である。最後の2つの次元を平坦化することによって次元(20,200)を有するように行列を形状変更し、20個のアミノ酸の各々に対して作用する重みが長さ200のベクトルとして表されるような行列を得た。20個のアミノ酸間の相関行列を計算した。各次元が各アミノ酸を表すので、相関係数行列を計算することによって、深層学習ネットワークが訓練データから学習したことに基づいて、アミノ酸間の相関と、深層学習ネットワークに対してアミノ酸がどれだけ類似しているように見えるかということとを計算している。相関係数行列の視覚化が図27に示されており(アミノ酸はBLOSUM62の行列順序でソートされている)、疎水性アミノ酸(メチオニン、イソロイシン、ロイシン、バリン、フェニルアラニン、チロシン、トリプトファン)および親水性アミノ酸(アスパラギン、アスパラギン酸、グルタミン酸、グルタミン、アルギニン、およびリジン)を備える2つの顕著なクラスタを示す。これらの初期層の出力はより後の層の入力になり、深層学習ネットワークがデータのますます複雑な階層的表現を構築することを可能にする。

【０５７７】

ニューラルネットワークが予測において使用するアミノ酸配列のウィンドウを示すために、5000個のランダムに選択されたバリアント内のおよびその周辺の各場所を摂動させて、バリアントに対する予測されるPrimateAIスコアに対する影響を観察した(図25B)。バリアントの周りの各々の近くのアミノ酸場所(-25～+25)における入力を系統的に0にして、

ニューラルネットワークにより予測されるバリアントの病原性の変化を測定し、5000個のバリアントにわたってその変化の平均絶対値をプロットした。バリアントの近くのアミノ酸が最も大きい影響を受けており、概ね対称的な分布で、バリアントからの距離が長くなるにつれて徐々に影響が小さくなる。重要なことに、このモデルは、バリアントの場所におけるアミノ酸だけに基づくのではなく、タンパク質モチーフを認識するために必要とされるであろうより広いウィンドウからの情報を使用することによって、予測を行う。タンパク質サブドメインが比較的小型のサイズであることと一致して、51個を超えるアミノ酸へとウィンドウのサイズを延長することが、さらに正確さを改善しないことを経験的に観察した。

【０５７８】

深層学習分類器のアラインメントに対する感度を評価するために、バリアント分類の正確さに対するアラインメントの深さの影響を次のように調査した。アラインメントにおける種の数に基づいてデータを5つのビンへと分割し、各ビンにおけるネットワークの正確さを評価した(図57)。トリヌクレオチドコンテクストに対して照合された(図21Dのように、しかし各ビンに対して別々に実行される)ランダムに選択された変異から、保留された良性変異のセットを分離する際のネットワークの正確さは、上位の3つのビンにおいて最も高く、下位の2つのビンにおいて顕著に弱いことを発見した。99種の脊椎動物の多種アラインメントは、11種のヒト以外の霊長類、50種の哺乳類、および38種の脊椎動物を備え、下位の2つのビンは、他の非霊長類の哺乳類からのまばらなアラインメント情報を有するタンパク質を表す。深層学習ネットワークは、アラインメント情報が霊長類および哺乳類全体に及ぶときにロバストかつ正確であり、より遠縁の脊椎動物からの保存情報はより重要性が低い。

【０５７９】

［正規のコーディング領域の定義］

正規のコーディング領域を定義するために、コーディングDNA配列(CDS)領域(knownCanonical.exonNuc.fa.gz)に対するヒトとの99種の脊椎動物ゲノムの複数アラインメントがUCSCゲノムブラウザからダウンロードされた。ヒトについては、エクソンの座標はBuild hg19のもとにある。エクソンは統合されて遺伝子を形成する。常染色体上の遺伝子およびchrXが保持される。相同ではない遺伝子は除去され、相同な遺伝子のリストはNCBI ftp://ftp.ncbi.nih.gov/pub/HomoloGene/current/homologene.dataからダウンロードされた。複数の遺伝子アノテーションを伴うSNPに対しては、SNPのアノテーションを表すために最長の転写産物が選択される。

【０５８０】

［ヒト、類人猿、および哺乳類の多型データ］

世界中の8つの亜集団からの123136人の個人の全エクソンシーケンシングデータを収集した、最近の大規模な研究であるgenome Aggregation Database(gnomAD)から、ヒトエクソン多型データをダウンロードした。そして、フィルタを通過し正規のコーディング領域に該当するバリアントを抽出した。

【０５８１】

大型類人猿ゲノムシーケンシングプロジェクトは、24体のチンパンジー、13体のボノボ、27体のゴリラ、および10体のオランウータン(5体のスマトラオランウータンおよび5体のボルネオオランウータンを含む)の全ゲノムシーケンシングデータを提供する。チンパンジーおよびボノボについての研究は、追加の25体の類人猿のWGSを提供する。すべてのシーケンシングデータはhg19にマッピングされたので、これらの研究からVCFファイルをダウンロードし、正規のコーディング領域内でバリアントを直接抽出した。

【０５８２】

他の類人猿および哺乳類と比較するために、アカゲザル、マーモセット、ブタ、ウシ、ヤギ、ネズミ、およびニワトリを含む少数の他の種のSNPもdbSNPからダウンロードした。イヌ、ネコ、またはヒツジなどの他の種は廃棄した。それは、dbSNPがそれらの種に対して限られた数のバリアントを提供するからである。最初に、各種のSNPをhg19にリフトオ

ーバーした。バリアントの約20%が偽遺伝子領域にマッピングされることが判明した。次いで、正規のコーディング領域の100種の脊椎動物の複数のアラインメントファイルから各種のエクソン座標を取得し、それらのエクソン内のバリアントを抽出した。次いで、それらの抽出されたSNPはhg19にリフトオーバーされた。バリアントがアラインメントとは異なる種のゲノムビルド上にある場合、まずアラインメントのゲノムビルドにバリアントをリフトした。

【０５８３】

ウシSNPデータは様々な研究に由来するので、ウシバリアントのすべての大きなバッチをdbSNPからダウンロードし(VCFファイルが100MBより大きい16個のバッチ)、各バッチに対するミスセンス対同義比を計算することによってウシSNPの様々なバッチの品質を評価した。ミスセンス対同義比の中央値は0.781であり、中央絶対偏差(MAD)は0.160である(平均は0.879でありSDは0.496である)。異常値の比を伴う2つのバッチ(比が1.391であるsnpBatch_1000_BULL_GENOMES_1059190.gzおよび比が2.568であるsnpBatch_COFACTOR_GENOMICS_1059634.gz)はさらなる分析から除外された。

【０５８４】

［類人猿および哺乳類における多型の性質の評価］

大型類人猿SNPの有用性を実証するために、シングルトンSNPと一般的なSNP(アレル頻度(AF)>0.1%)の数の比を測定するエンリッチメントスコアを考案した。同義バリアントは、良性でありどのような選択圧力も受けずに一般に中立的に進化することが知られている。有害なミスセンスバリアントは、自然選択によって徐々に排除されるので、そのアレル頻度分布は同義バリアントと比較して稀なバリアントが多い傾向がある。

【０５８５】

霊長類、哺乳類、および家禽において観察されるSNPと重複するgnomAD SNPに注目した。種毎の同義バリアントおよびミスセンスバリアントの数をカウントした。ミスセンスバリアントについては、「ミスセンス同一(missense identical)」と名付けられる、別の種において同一のアミノ酸変化を共有するタイプと、「ミスセンス相違(missense different)」と名付けられる、別の種において異なるアミノ酸変化を有するタイプという、2つのタイプへとさらに分類した。次いで、シングルトンバリアントの数と一般的なバリアントの数の比として、エンリッチメントスコアが種毎に計算された。

【０５８６】

加えて、各種について同義バリアントとミスセンス同一バリアントとの間でエンリッチメントスコアを比較するために、2×2の分割表に対して相同性のカイ二乗($\chi 2$)検定を実行した。すべての霊長類が、同義バリアントとミスセンス同一バリアントとの間でエンリッチメントスコアに有意な差を示さないが、ウシ、ネズミ、およびニワトリは有意な差を示す。

【０５８７】

この結果は、大型類人猿において同一のアミノ酸変化を共有するSNPが、同義SNPと非常に類似するエンリッチメントスコアを有する傾向があることを明らかにしており、それらのSNPに、ヒトの健康に対する軽度の影響があることを示唆している。異なるアミノ酸変化を有する、または大型類人猿において存在しないSNPは、同義SNPのエンリッチメントスコアから有意に逸脱するエンリッチメントスコアを有するが、非霊長類の種からのミスセンス多型も、同義バリアントと異なるアレル頻度分布を有する。結論は、大型類人猿において同一のアミノ酸変化を共有するSNPを、良性バリアントの訓練セットに追加することができるということである。

【０５８８】

我々の仮定は、大半のバリアントが独立に派生したものであり、家系同一性(IBD)により生成されるのではないということである。したがって、IBD SNPにおける稀なバリアントのエンリッチメント分析を実行して、それらのエンリッチメントスコアの様々な挙動を評価した。IBD SNPは、チンパンジー、ボノボ、ゴリラ、S.オランウータン、およびB.オランウータンを含む、2つ以上の大型類人猿の種とヒトとの両方において現れる、ヒトSNP

として定義される。次いで、シングルトンの数を一般的なバリアント(AF>0.1%)の数で割ったものとして定義されるエンリッチメントスコアが、ミスセンスバリアントおよび同義バリアントに対して別々に計算され、同義バリアントは中立的であると考えられ比較の基準として働く。

【０５８９】

［哺乳類の種間での固定された置換］

固定された置換のエンリッチメント分析

種間の置換の稀なバリアントエンリッチメント分析も研究した。UCSCゲノムブラウザ(http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/hg19.100way.commonNames.nh)から100種の脊椎動物の種の進化系統樹をダウンロードした。次いで、計算されたペア毎の進化系統的距離を計算し、近縁の種のペア(距離<0.3)を選択した。霊長類の種のペアを得るために、UCSCゲノムブラウザからCDS領域に対するヒトとの19種の哺乳類(16種の霊長類)ゲノムのアラインメント(hg38)をダウンロードした。4つの霊長類のペアが13個の脊椎動物のペアに追加された。以下の表は、一実装形態による、近縁の種の複数のペアの遺伝的距離を示す。

【０５９０】

【表８】

種１	種２	距離
チンパンジー	ボノボ	0.00799195
アカゲザル	カニクイザル	0.00576807
マーモセット	リスザル	0.0680206
テングザル	キンシコウ	0.01339514
ヒヒ	ミドリザル	0.045652
ウマ	シロサイ	0.084397
ヒツジ	家畜のヤギ	0.1
マウス	ラット	0.176098
チンチラ	フサオネズミ	0.16
ケープキンモグラ	テンレック	0.265936
チャイニーズハムスター	ゴールデンハムスター	0.08
ホオヒゲコウモリ	ココウモリ	0.08254
イルカ	シャチ	0.074368
セーカーハヤブサ	ハヤブサ	0.2
ニワトリ	シチメンチョウ	0.126972
アオウミガメ	ニシキガメ	0.2
スッポン	トゲスッポン	0.2

【０５９１】

正規のコーディング領域内でヒトとの19種の哺乳類または99種の脊椎動物のゲノムの複数アラインメントを取り込み、脊椎動物の各々の選択されたペア間のヌクレオチド置換を得た。これらの置換は、種のペアとヒトバリアントとの間のコドン変化が同一であることを条件として、gnomADからヒトエクソンSNPにマッピングされた。バリアントを、同義バ

リアント、別の種において同一のアミノ酸変化を共有するミスセンスバリアント、および別の種において異なるアミノ酸変化を有するミスセンスバリアントという、3つのタイプへと分類した。エンリッチメントスコアが種のペア毎に各クラスに対して計算された。

【０５９２】

種内多型および種間多型の比較

チンパンジー、アカゲザル、マーモセット、ヤギ、ネズミ、およびニワトリを含む6つの種が、種内多型および種間多型の比較を実行するために選択され、それは、これらの種については種内バリアントと種間バリアントの両方が利用可能であったからである。種内バリアントおよび種間バリアントのエンリッチメントスコアの比較は、2つの2×2の分割表のオッズ比の比較に類似している。通常は、分割表間のオッズ比の相同性を評価するために、Woolf検定が適用される。したがって、Woolf検定を利用して、種内多型と種間多型との間のエンリッチメントスコアの差を評価した。

【０５９３】

［遺伝子毎のエンリッチメント分析］

図64は、遺伝子毎のエンリッチメント分析の一実装形態を示す。一実装形態では、深層畳み込みニューラルネットワークベースのバリアント病原性分類器はさらに、病原性であると決定されたバリアントの病原性を確認する遺伝子毎のエンリッチメント分析を実施するように構成される。遺伝的疾患を持つ個人のコホートからサンプリングされた特定の遺伝子に対して、遺伝子毎のエンリッチメント分析は、病原性である特定の遺伝子におけるバリアント候補を特定するために深層畳み込みニューラルネットワークベースのバリアント病原性分類器を適用することと、バリアント候補の観察されるトリヌクレオチド変異率を合計してその合計を送信カウントおよびコホートのサイズと乗じることに基づいて特定の遺伝子に対する変異の基準数を決定することと、病原性である特定の遺伝子の中のde novoミスセンスバリアントを特定するために深層畳み込みニューラルネットワークベースのバリアント病原性分類器を適用することと、変異の基準数をde novoミスセンスバリアントのカウントと比較することとを含む。比較の出力に基づいて、遺伝子毎のエンリッチメント分析は、特定の遺伝子が遺伝子障害と関連付けられることと、de novoミスセンスバリアントが病原性であることとを確認する。いくつかの実装形態では、遺伝子障害は自閉スペクトラム障害(ASDと省略される)である。他の実装形態では、遺伝的障害は発達遅延障害(DDDと省略される)である。

【０５９４】

図64に示される例では、特定の遺伝子の中の5つのバリアント候補は、深層畳み込みニューラルネットワークベースのバリアント病原性分類器によって病原性であるものとして分類された。これらの5つのバリアント候補は、10^{-8}、10^{-2}、10^{-1}、10^{5}、および10^{1}という観察されたそれぞれのトリヌクレオチド変異率を有する。特定の遺伝子に対する変異の基準数は、5つのバリアント候補のそれぞれの観察されたトリヌクレオチド変異率を合計し、その合計を送信/染色体カウント(2)およびコホートのサイズ(1000)と乗じることに基づいて、10^{-5}であると決定される。これが次いでde novoバリアントカウント(3)と比較される。

【０５９５】

いくつかの実装形態では、深層畳み込みニューラルネットワークベースのバリアント病原性分類器はさらに、出力としてp値を生み出す統計的検定を使用して比較を実行するように構成される。

【０５９６】

他の実装形態では、深層畳み込みニューラルネットワークベースのバリアント病原性分類器はさらに、変異の基準数をde novoミスセンスバリアントのカウントと比較し、比較の出力に基づいて、特定の遺伝子が遺伝的疾患と関連付けられないことと、de novoミスセンスバリアントが良性であることとを確認するように構成される。

【０５９７】

［ゲノムワイドエンリッチメント分析］

図65は、ゲノムワイドエンリッチメント分析の一実装形態を示す。別の実装形態では、深層畳み込みニューラルネットワークベースのバリアント病原性分類器はさらに、病原性と決定されたバリアントの病原性を確認するゲノムワイドエンリッチメント分析を実施するように構成される。ゲノムワイドエンリッチメント分析は、健康な個人のコホートからサンプリングされた複数の遺伝子において病原性であるde novoミスセンスバリアントの第1のセットを特定するために深層畳み込みニューラルネットワークベースのバリアント病原性分類器を適用することと、遺伝子障害を持つ個人のコホートからサンプリングされる複数の遺伝子において病原性であるde novoミスセンスバリアントの第2のセットを特定するために深層畳み込みニューラルネットワークベースのバリアント病原性分類器を適用することと、第1のセットおよび第2のセットのそれぞれのカウントを比較することと、比較の出力に基づいて、de novoミスセンスバリアントの第2のセットが遺伝的障害を持つ個人のコホートにおいてエンリッチされ、したがって病原性であることを確認することとを含む。いくつかの実装形態では、遺伝的疾患は自閉スペクトラム障害(ASDと省略される)である。他の実装形態では、遺伝的障害は発達遅延障害(DDDと省略される)である。

【０５９８】

いくつかの実装形態では、深層畳み込みニューラルネットワークベースのバリアント病原性分類器はさらに、p値を出力として生み出す統計的検定を使用して比較を実行するように構成される。一実装形態では、比較はさらにそれぞれのコホートサイズによってパラメータ化される。

【０５９９】

いくつかの実装形態では、深層畳み込みニューラルネットワークベースのバリアント病原性分類器はさらに、第1のセットおよび第2のセットのそれぞれのカウントを比較し、比較の出力に基づいて、de novoミスセンスバリアントの第2のセットが遺伝的障害を持つ個人のコホートにおいてエンリッチされず、したがって良性であることを確認するように構成される。

【０６００】

図65に示される例では、健康なコホートにおける変異率(0.001)および影響を受けているコホートにおける変異率(0.004)が、個人毎の変異率(4)とともに示されている。

【０６０１】

［具体的な実装形態］

バリアント病原性分類器を構築するためのシステム、方法、および製造物品を説明する。実装形態の1つまたは複数の特徴は基本の実装形態と組み合わされ得る。相互に排他的ではない実装形態は合成可能であると教示される。実装形態の1つまたは複数の特徴は他の実装形態と合成され得る。本開示は定期的にこれらの選択肢をユーザに思い起こさせる。これらの選択肢を繰り返し述べる記載がいくつかの実装形態において省略されていることは、先行するセクションにおいて教示された合成を限定するものと解釈されるべきではなく、これらの記載は以後の実装形態の各々へと前方に参照によって組み込まれる。

【０６０２】

開示される技術のシステム実装形態は、メモリに結合される1つまたは複数のプロセッサを含む。メモリは、ゲノム配列(たとえば、ヌクレオチド配列)におけるスプライスサイトを特定するスプライスサイト検出器を訓練するためのコンピュータ命令をロードされる。

【０６０３】

図48および図19に示されるように、システムは畳み込みニューラルネットワークベースのバリアント病原性分類器を訓練し、これはメモリに結合される多数のプロセッサ上で実行される。システムは、良性バリアントおよび病原性バリアントから生成されたタンパク質配列ペアの、良性訓練例および病原性訓練例を使用する。良性バリアントは、一般的なヒトミスセンスバリアントと、ヒトと一致する基準コドン配列を共有する代替のヒト以外の霊長類コドン配列上で発生するヒト以外の霊長類ミスセンスバリアントとを含む。「タンパク質配列ペア」という語句は、基準タンパク質配列および代替タンパク質配列を指し

、基準タンパク質配列は基準トリプレットヌクレオチド塩基(基準コドン)によって形成される基準アミノ酸を備え、代替タンパク質配列は代替トリプレットヌクレオチド塩基(代替コドン)によって形成される代替アミノ酸を備えるので、代替タンパク質配列は基準タンパク質配列の基準アミノ酸を形成する基準トリプレットヌクレオチド塩基(基準コドン)において発生するバリアントの結果として作り出される。バリアントは、SNP、挿入、または欠失であり得る。

【０６０４】

このシステムの実装形態および開示される他のシステムは任意選択で、以下の特徴のうちの1つまたは複数を含む。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。簡潔にするために、システム特徴の代替的な組合せは個別に列挙されない。システム、方法、製造物品に適用可能な特徴は、基本の特徴の各々のstatutory classのセットに対して繰り返されない。このセクションにおいて特定される特徴がどのように他のstatutory classの中の基本の特徴と容易に組み合わされ得るかを、読者は理解するであろう。

【０６０５】

図44に示されるように、一般的なヒトミスセンスバリアントは、少なくとも100000人のヒトからサンプリングされたヒト集団バリアントデータセットにわたって0.1%より高いマイナーアレル頻度(MAFと省略される)を有する。

【０６０６】

図44に示されるように、サンプリングされたヒトは異なるヒト亜集団に属し、一般的なヒトミスセンスバリアントはそれぞれのヒト亜集団バリアントデータセット内で0.1%より高いMAFを有する。

【０６０７】

ヒト亜集団は、アフリカ人/アフリカ系アメリカ人(AFRと省略される)、アメリカ人(AMRと省略される)、アシュケナージ系ユダヤ人(ASJと省略される)、東アジア人(EASと省略される)、フィンランド人(FINと省略される)、フィンランド人以外のヨーロッパ人(NFEと省略される)、南アジア人(SASと省略される)、および他(OTHと省略される)を含む。

【０６０８】

図43および図44に示されるように、ヒト以外の霊長類ミスセンスバリアントは、チンパンジー、ボノボ、ゴリラ、B.オランウータン、S.オランウータン、アカゲザル、およびマーモセットを含む、複数のヒト以外の霊長類の種からのミスセンスバリアントを含む。

【０６０９】

図45および図46に示されるように、エンリッチメント分析に基づいて、システムは、ある特定のヒト以外の霊長類の種を、良性バリアントにその特定のヒト以外の霊長類の種のミスセンスバリアントを含めるために受け入れる。エンリッチメント分析は、特定のヒト以外の霊長類の種に対して、特定のヒト以外の霊長類の種の同義バリアントの第1のエンリッチメントスコアを、特定のヒト以外の霊長類の種のミスセンス同一バリアントの第2のエンリッチメントスコアと比較することを含む。

【０６１０】

図45は、ヒトオーソログガスミスセンスSNPの一実装形態を示す。ヒト以外の種におけるミスセンスSNPは、ヒトと一致する基準コドンおよび代替コドンを有する。図45に示されるように、ミスセンス同一バリアントは、ヒトと一致する基準コドン配列および代替コドン配列を共有するミスセンスバリアントである。

【０６１１】

図46および図47に示されるように、第1のエンリッチメントスコアは、MAFが0.1%より大きい一般的な同義バリアントに対する、MAFが0.1%より小さい稀な同義バリアントの比を決定することによって、作り出される。第2のエンリッチメントスコアは、MAFが0.1%より大きい一般的なミスセンス同一バリアントに対する、MAFが0.1%より小さい稀なミスセンス同一バリアントの比を決定することによって、作り出される。稀なバリアントは、シングルトンバリアントを含む。

【０６１２】
図46および図47に示されるように、第1のエンリッチメントスコアと第2のエンリッチメントスコアの差は所定の範囲内にあり、良性バリアントに特定のヒト以外の霊長類のミスセンスバリアントを含めるために、その特定のヒト以外の霊長類の種を受け入れることをさらに含む。差が所定の範囲にあることは、ミスセンス同一バリアントが同義バリアントと同じ程度の自然選択を受けているので、同義バリアントと同じくらい良性であることを示す。
【０６１３】
図48に示されるように、システムは、良性バリアントにヒト以外の霊長類の種のミスセンスバリアントを含めるために、複数のヒト以外の霊長類の種を受け入れるようにエンリッチメント分析を繰り返し適用する。システムはさらに、ヒト以外の霊長類の種の各々に対する同義バリアントの第1のエンリッチメントスコアとミスセンス同一バリアントの第2のエンリッチメントスコアを比較するための、相同性のカイ二乗検定を含む。
【０６１４】
図48に示されるように、ヒト以外の霊長類ミスセンスバリアントのカウントは少なくとも100000である。ヒト以外の霊長類ミスセンスバリアントのカウントは385236である。一般的なヒトミスセンスバリアントのカウントは少なくとも50000である。一般的なヒトミスセンスバリアントのカウントは83546である。
【０６１５】
他の実装形態は、上で説明されたシステムの活動を実行するようにプロセッサによって実行可能な命令を記憶する、非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、上で説明されたシステムの活動を実行する方法を含み得る。
【０６１６】
開示される技術の別のシステムの実装形態は、一塩基多型(SNPと省略される)病原性分類器を構築することを含む。システムは畳み込みニューラルネットワークベースのSNP病原性分類器を訓練し、これは、良性SNPおよび病原性SNPによって表されるアミノ酸配列の良性訓練例および病原性訓練例を使用して、メモリに結合された多数のプロセッサ上で実行される。良性訓練例は、アミノ酸配列ペアとして表されるヌクレオチド配列の第1のセットおよび第2のセットを含み、各アミノ酸配列は、上流および下流のアミノ酸が側にある中心アミノ酸を含む。各アミノ酸配列ペアは、基準ヌクレオチド配列によって表されるアミノ酸の基準配列と、SNPを含む代替ヌクレオチド配列によって表されるアミノ酸の代替配列とを含む。
【０６１７】
図9に示されるように、第1のセットはヒトヌクレオチド配列ペアを備え、各ペアは、SNPを含みヒト集団内で一般的であると見なされるマイナーアレル頻度(MAFと省略される)を有するヒト代替ヌクレオチド配列を含む。第2のセットは、ヒト以外の霊長類代替ヌクレオチド配列とペアにされたヒト以外の霊長類基準ヌクレオチド配列を備える。ヒト以外の霊長類基準ヌクレオチド配列は、オーソログスなヒトヌクレオチド基準配列を有する。ヒト以外の霊長類代替ヌクレオチド配列はSNPを含む。
【０６１８】
第1のシステムの実装形態についてこの特定の実装セクションにおいて論じられる特徴の各々は、このシステムの実装形態に等しく適用される。上で示されたように、すべてのシステム特徴はここで繰り返されず、参照によって繰り返されるものと見なされるべきである。
【０６１９】
他の実装形態は、上で説明されたシステムの活動を実行するようにプロセッサによって実行可能な命令を記憶する、非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、上で説明されたシステムの活動を実行する方法を含み得る。
【０６２０】
図48および図19に示されるように、開示される技術の第1の方法の実装形態は、バリア

ント病原性分類器を構築するステップを含む。方法はさらに、畳み込みニューラルネットワークベースのバリアント病原性分類器を訓練するステップを含み、これは、良性バリアントおよび病原性バリアントから生成されるタンパク質配列ペアの良性訓練例および病原性訓練例を使用して、メモリに結合される多数のプロセッサ上で実行される。良性バリアントは、一般的なヒトミスセンスバリアントと、ヒトと一致する基準コドン配列を共有する代替的なヒト以外の霊長類コドン配列上で発生するヒト以外の霊長類ミスセンスバリアントとを含む。

【０６２１】

第1のシステム実装形態についてこの特定の実装セクションにおいて論じられる特徴の各々は、この方法の実装形態に等しく適用される。上で示されたように、すべてのシステム特徴はここで繰り返されず、参照によって繰り返されるものと見なされるべきである。

【０６２２】

他の実装形態は、上で説明された方法を実行するようにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、メモリと、上で説明された方法を実行するためにメモリに記憶された命令を実行するように動作可能な1つまたは複数のプロセッサとを含む、システムを含み得る。

【０６２３】

図48および図19に示されるように、開示される技術の第2の方法の実装形態は、一塩基多型(SNPと省略される)病原性分類器を構築するステップを含む。方法はさらに、畳み込みニューラルネットワークベースのSNP病原性分類器を訓練するステップを含み、これは、良性SNPおよび病原性SNPによって表されるアミノ酸配列の良性訓練例および病原性訓練例を使用して、メモリに結合される多数のプロセッサ上で実行される。良性訓練例は、アミノ酸配列のペアとして表されるヌクレオチド配列の第1のセットおよび第2のセットを含み、各アミノ酸配列は上流および下流のアミノ酸が側にある中心アミノ酸を含み、各アミノ酸配列のペアは、基準ヌクレオチド配列によって表されるアミノ酸の基準配列およびSNPを含む代替ヌクレオチド配列によって表されるアミノ酸の代替配列を含む。第1のセットはヒトヌクレオチド配列ペアを備え、各ペアは、SNPを含みヒト集団内で一般的であると見なされるマイナーアレル頻度(MAFと省略される)を有する、ヒト代替ヌクレオチド配列を含む。第2のセットは、ヒト以外の霊長類代替ヌクレオチド配列とペアにされた、ヒト以外の霊長類基準ヌクレオチド配列を備える。ヒト以外の霊長類基準ヌクレオチド配列はオーソロガスなヒトヌクレオチド基準配列を有し、ヒト以外の霊長類代替ヌクレオチド配列はSNPを含む。

【０６２４】

第2のシステム実装形態についてこの特定の実装セクションにおいて論じられる特徴の各々は、この方法の実装形態に等しく適用される。上で示されたように、すべてのシステム特徴はここで繰り返されず、参照によって繰り返されるものと見なされるべきである。

【０６２５】

他の実装形態は、上で説明された方法を実行するようにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、メモリと、上で説明された方法を実行するためにメモリに記憶された命令を実行するように動作可能な1つまたは複数のプロセッサとを含む、システムを含み得る。

【０６２６】

二次構造および溶媒接触性分類器を伴う深層畳み込みニューラルネットワークベースのバリアント病原性分類器を使用するためのシステム、方法、および製造物品を説明する。実装形態の1つまたは複数の特徴は基本の実装形態と合成され得る。相互に排他的ではない実装形態は、合成可能であると教示される。実装形態の1つまたは複数の特徴は他の実装形態と合成され得る。本開示は定期的にこれらの選択肢をユーザに思い起こさせる。これらの選択肢を繰り返し述べる記載がいくつかの実装形態において省略されていることは、先行するセクションにおいて教示された合成を限定するものと解釈されるべきではなく、これらの記載は以後の実装形態の各々へと前方に参照によって組み込まれる。

【０６２７】
開示される技術のシステム実装形態は、メモリに結合される1つまたは複数のプロセッサを含む。メモリは、二次構造および溶媒接触性分類器を伴う深層畳み込みニューラルネットワークベースのバリアント病原性分類器を実行するためのコンピュータ命令をロードされる。
【０６２８】
システムは、タンパク質配列内のアミノ酸位置において3状態二次構造を予測するように訓練される、メモリに結合された多数のプロセッサ上で実行される第1の二次構造サブネットワークを備える。システムはさらに、タンパク質配列内のアミノ酸位置において3状態溶媒接触性を予測するように訓練される、メモリに結合された多数のプロセッサ上で実行される第2の溶媒接触性サブネットワークを備える。
【０６２９】
3状態二次構造は、複数のDNA二次構造状態であるαヘリックス(H)、βシート(B)、およびコイル(C)のうちの1つを指す。
【０６３０】
3状態溶媒接触性は、複数のタンパク質溶媒接触性である埋もれている(B)、中間(I)、および露出している(E)のうちの1つを指す。
【０６３１】
多数のプロセッサのうちの少なくとも1つで実行される位置特定的頻度行列(PFMと省略される)生成器は、霊長類PFM、哺乳類PFM、および脊椎動物PFMを生成するために、霊長類、哺乳類、および霊長類と哺乳類を除く脊椎動物の3つの配列グループに適用される。
【０６３２】
言い換えると、これは、PFM生成器を霊長類配列データに適用して霊長類PFMを生成することと、PFM生成器を哺乳類配列データに適用して哺乳類PFMを生成することと、PFM生成器を霊長類配列データおよび哺乳類配列データを含まない脊椎動物配列データに適用して脊椎動物PFMを生成することとを含む。
【０６３３】
入力プロセッサは、標的バリアントアミノ酸の上流および下流の側に各方向への少なくとも25個のアミノ酸があるバリアントアミノ酸配列を受け入れ、一塩基バリアントが標的バリアントアミノ酸を生み出す。多数のプロセッサのうちの少なくとも1つで実行される補足データ割振器は、そのバリアントアミノ酸配列とアラインメントされた、標的基準アミノ酸の上流および下流の側に各方向への少なくとも25個のアミノ酸がある基準アミノ酸配列を割り振る。これに続いて、補足データ割振器は、基準アミノ酸配列のために第1のサブネットワークおよび第2のサブネットワークによって作り出される基準状態分類を割り振る。この後で、補足データ割振器は、バリアントアミノ酸配列のために第1のサブネットワークおよび第2のサブネットワークによって作り出されるバリアント状態分類を割り振る。最後に、補足データ割振器は、基準アミノ酸配列とアラインメントされる霊長類PFM、哺乳類PFM、および脊椎動物PFMを割り振る。
【０６３４】
本出願の文脈では、「とアラインメントされる」という語句は、基準アミノ酸配列または代替アミノ酸配列の中の各アミノ場所に対する霊長類PFM、哺乳類PFM、および脊椎動物PFMを場所毎に決定し、アミノ酸場所が基準アミノ酸配列または代替アミノ酸配列において発生するのと同じ順序で場所毎にまたは場所の序数に基づいてその決定の結果を符号化して記憶することを指す。
【０６３５】
システムはまた、バリアントアミノ酸配列、割り振られた基準アミノ酸配列、割り振られた基準状態分類およびバリアント状態分類、ならびに割り振られたPFMを処理したことに基づいて、良性または病原性であるものとしてバリアントアミノ酸配列を分類するように訓練された、多数のプロセッサで実行される深層畳み込みニューラルネットワークを含む。システムは、バリアントアミノ酸配列に対する病原性スコアを少なくとも報告する出

力プロセッサを含む。

【０６３６】

このシステム実装形態および開示される他のシステムは任意選択で、以下の特徴のうちの1つまたは複数を含む。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。簡潔にするために、システム特徴の代替的な組合せは個別に列挙されない。システム、方法、および製造物品に適用可能な特徴は、基本の特徴の各statutory classセットに対して繰り返されない。読者は、このセクションにおいて特定される特徴が他のstatutory classにおける基本の特徴とどのように容易に合成され得るかを理解するであろう。

【０６３７】

深層畳み込みニューラルネットワークベースのバリアント病原性分類器を備えるシステムはさらに、病原性スコアに基づいて良性または病原性として一塩基バリアントを分類するように構成される。

【０６３８】

システムは、深層畳み込みニューラルネットワークが、少なくとも、バリアントアミノ酸配列、割り振られた基準アミノ酸配列、割り振られたバリアント二次構造状態分類、割り振られた基準二次構造状態分類、割り振られたバリアント溶媒接触性状態分類、割り振られた基準溶媒接触性状態分類、割り振られた霊長類PFM、割り振られた哺乳類PFM、および割り振られた脊椎動物PFMを、入力として並列に受け入れるような、深層畳み込みニューラルネットワークベースのバリアント病原性分類器を備える。

【０６３９】

システムは、バッチ正規化層、ReLU非線形性層、および次元変更層を使用して、バリアントアミノ酸配列、割り振られた基準アミノ酸配列、割り振られた霊長類PFM、割り振られた哺乳類PFM、および割り振られた脊椎動物PFMを前処理するように構成される。システムはさらに、前処理された特性評価を合計し、その合計を、割り振られたバリアント二次構造状態分類、割り振られた基準二次構造状態分類、割り振られたバリアント溶媒接触性状態分類、および割り振られた基準溶媒接触性状態分類と連結して、連結された入力を生み出すように構成される。システムは、次元変更層を通じて連結された入力を処理し、処理された連結された入力を受け入れて深層畳み込みニューラルネットワークの残差ブロックを開始する。

【０６４０】

深層畳み込みニューラルネットワークは、配列において最低から最高まで並べられる残差ブロックのグループを備える。深層畳み込みニューラルネットワークは、残差ブロックの数、スキップ接続の数、および非線形活性化を伴わない残差接続の数によってパラメータ化される。深層畳み込みニューラルネットワークは、先行する入力の空間次元および特徴量次元を形状変更する次元変更層を備える。

【０６４１】

システムはさらに、霊長類、哺乳類、および脊椎動物にわたってアライメントされた基準アミノ酸配列において保存されている標的基準アミノ酸から標的バリアントアミノ酸を作り出す、一塩基バリアントを病原性として分類するように訓練するように構成される。

【０６４２】

保存率は、標的基準アミノ酸の機能的な有意性を表し、PFWから決定される。システムはさらに、バリアントアミノ酸配列と基準バリアントアミノ酸配列との間で異なる二次構造を引き起こす一塩基バリアントを病原性として分類するように訓練するように構成される。

【０６４３】

システムはさらに、バリアントアミノ酸配列と基準バリアントアミノ酸配列との間で異なる溶媒接触性を引き起こす一塩基バリアントを病原性として分類するように訓練するように構成される。

【０６４４】

PFMは、他の種のアライメントされるタンパク質配列にわたるヒトタンパク質配列におけるアミノ酸の出現の頻度を位置ごとに決定することによって、他の種のアライメントされるタンパク質配列にわたるヒトタンパク質配列におけるアミノ酸の保存率を表す。

【０６４５】

二次構造の3状態は、ヘリックス、シート、およびコイルである。第1の二次構造サブネットワークは、入力タンパク質配列と、入力タンパク質配列内のアミノ酸位置とアライメントされる霊長類PFM、哺乳類PFM、および脊椎動物PFMとを受け入れ、アミノ酸位置の各々において3状態二次構造を予測するように訓練される。溶媒接触性の3状態は、露出している、埋もれている、および中間である。

【０６４６】

二次溶媒接触性サブネットワークは、入力タンパク質配列と、入力タンパク質配列内のアミノ酸位置とアラインメントされている霊長類PFM、哺乳類PFM、および脊椎動物PFMとを受け入れ、アミノ酸位置の各々において3状態溶媒接触性を予測するように訓練される。入力タンパク質配列は基準タンパク質配列である。入力タンパク質配列は代替タンパク質配列である。第1の二次構造サブネットワークは、配列において最低から最高まで並べられる残差ブロックのグループを備える。第1の二次構造サブネットワークは、残差ブロックの数、スキップ接続の数、および非線形活性化を伴わない残差接続の数によってパラメータ化される。

【０６４７】

第1の二次構造サブネットワークは、先行する入力の空間次元および特徴量次元を形状変更する次元変更層を備える。第2の溶媒接触性サブネットワークは、配列において最低から最高まで並べられる残差ブロックのグループを備える。第2の溶媒接触性サブネットワークは、残差ブロックの数、スキップ接続の数、および非線形活性化を伴わない残差接続の数によってパラメータ化される。第2の溶媒接触性サブネットワークは、先行する入力の空間次元および特徴量次元を形状変更する次元変更層を備える。

【０６４８】

各残差ブロックは、少なくとも1つのバッチ正規化層、少なくとも1つの正規化線形ユニット(ReLUと省略される)層、少なくとも1つの次元変更層、および少なくとも1つの残差接続を備える。各残差ブロックは、2つのバッチ正規化層、2つのReLU非線形性層、2つの次元変更層、および1つの残差接続を備える。

【０６４９】

深層畳み込みニューラルネットワーク、第1の二次構造サブネットワーク、および第2の溶媒接触性サブネットワークは各々、最終分類層を備える。最終分類層はシグモイドベースの層である。最終分類層はソフトマックスベースの層である。

【０６５０】

システムはさらに、深層畳み込みニューラルネットワークとの協調のために、第1の二次構造サブネットワークおよび第2の溶媒接触性サブネットワークの最終分類層を除去するように構成される。

【０６５１】

システムはさらに、深層畳み込みニューラルネットワークの訓練の間に、誤差をサブネットワークに逆伝播してサブネットワーク重みを更新することを含めて、第1の二次構造サブネットワークおよび第2の溶媒接触性サブネットワークを病原性分類についてさらに訓練するように構成される。

【０６５２】

第2の溶媒接触性サブネットワークは少なくとも膨張畳み込み層を備える。システムはさらに、発達遅延障害(DDDと省略される)を引き起こすバリアントを病原性として分類するように構成される。バリアントアミノ酸配列および基準アミノ酸配列はフランキングアミノ酸を共有する。システムはさらに、深層畳み込みニューラルネットワークへの入力を符号化するためにワンホット符号化を使用するように構成される。

【０６５３】

図1Qは、開示される技術が動作することのできる例示的なコンピューティング環境を示す。深層畳み込みニューラルネットワーク、第1の二次構造サブネットワーク、および第2の溶媒接触性サブネットワークは、1つまたは複数の訓練サーバ上で訓練される。訓練された深層畳み込みニューラルネットワーク、第1の訓練された二次構造サブネットワーク、および訓練された第2の溶媒接触性サブネットワークは、要求側のクライアントから入力配列を受け取る1つまたは複数の本番サーバ上に展開される。本番サーバは、深層畳み込みニューラルネットワーク、第1の二次構造サブネットワーク、および第2の溶媒接触性サブネットワークのうちの少なくとも1つを通じて入力配列を処理して、クライアントに送信される出力を作り出す。

【０６５４】

他の実装形態は、上で説明されたシステムの活動を実行するようにプロセッサによって実行可能な命令を記憶する、非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、上で説明されたシステムの活動を実行する方法を含み得る。

【０６５５】

開示される技術の別のシステム実装形態は、メモリに結合された多数のプロセッサ上で実行される、深層畳み込みニューラルネットワークベースのバリアント病原性分類器を含む。システムは、霊長類PFMおよび哺乳類PFMを生成するために霊長類および哺乳類の2つの配列グループに適用される、多数のプロセッサのうちの少なくとも1つで実行される、位置特定的頻度行列(PFMと省略される)生成器を含む。システムはまた、標的バリアントアミノ酸の上流および下流の側に各方向への少なくとも25個のアミノ酸があるバリアントアミノ酸配列を受け入れる入力プロセッサを含み、一塩基バリアントが標的バリアントアミノ酸を生み出す。システムはまた、バリアントアミノ酸配列とアラインメントされる、標的基準アミノ酸の上流および下流の側に各方向への少なくとも25個のアミノ酸がある基準アミノ酸配列を割り振る、多数のプロセッサのうちの少なくとも1つで実行される補足データ割振器を含む。補足データ割振器はまた、基準アミノ酸配列とアラインメントされた霊長類PFMおよび哺乳類PFMを割り振る。システムはさらに、バリアントアミノ酸配列、割り振られた基準アミノ酸配列、および割り振られたPFMを処理することに基づいてバリアントアミノ酸配列を良性または病原性として分類するように訓練される、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を含む。最後に、システムは、バリアントアミノ酸配列に対する病原性スコアを少なくとも報告する出力プロセッサを含む。

【０６５６】

このシステム実装形態および開示される他のシステムは任意選択で、以下の特徴のうちの1つまたは複数を含む。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。簡潔にするために、システム特徴の代替的な組合せは個別に列挙されない。システム、方法、および製造物品に適用可能な特徴は、基本の特徴の各statutory classに対して繰り返されない。このセクションにおいて特定される特徴が他のstatutory classの中の基本の特徴とどのように容易に組み合わされ得るかを、読者は理解するであろう。

【０６５７】

システムはさらに、病原性スコアに基づいて一塩基バリアントを良性または病原性として分類するように構成される。深層畳み込みニューラルネットワークは、バリアントアミノ酸配列、割り振られた基準アミノ酸配列、割り振られた霊長類PFM、および割り振られた哺乳類PFMを並列に受け入れて処理する。システムはさらに、霊長類および哺乳類にわたって基準アミノ酸配列において保存されている標的基準アミノ酸から標的バリアントアミノ酸を作り出す、一塩基バリアントを病原性として分類するように訓練するように構成される。保存率は、標的基準アミノ酸の機能的な有意性を表し、PFWから決定される。

【０６５８】

第1のシステム実装形態に対してこの特定の実装セクションにおいて論じられる特徴の各々はこのシステム実装形態に等しく適用される。上で示されたように、すべてのシステム特徴は、ここでは繰り返されず、参照によって繰り返されるものと見なされるべきであ

る。

【０６５９】

他の実装形態は、上で説明されたシステムの活動を実行するようにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、上で説明されたシステムの活動を実行する方法を含み得る。

【０６６０】

開示される技術の第1の方法の実装形態は、タンパク質配列内のアミノ酸位置において3状態二次構造を予測するように訓練される、メモリに結合された多数のプロセッサ上で第1の二次構造サブネットワークを実行するステップを含む。タンパク質配列内のアミノ酸位置において3状態溶媒接触性を予測するように訓練される、メモリに結合される多数のプロセッサ上で第2の溶媒接触性サブネットワークを実行すること。多数のプロセッサのうちの少なくとも1つで実行すること。霊長類位置特定的頻度行列(PFMと省略される)、哺乳類PFM、および脊椎動物PFMを生成するために、霊長類、哺乳類、および霊長類と哺乳類を除く脊椎動物の3つの配列グループに適用される、PFM生成器。標的バリアントアミノ酸の上流および下流の側に各方向への少なくとも25個のアミノ酸があるバリアントアミノ酸配列入力プロセッサを受け入れること。一塩基バリアントは標的バリアントアミノ酸を作り出す。バリアントアミノ酸配列とアラインメントされた、標的基準アミノ酸の上流および下流の側に各方向への少なくとも25個のアミノ酸がある基準アミノ酸配列を割り振る補足データ割振器を、多数のプロセッサのうちの1つで実行すること。補足データ割振器はまた、基準アミノ酸配列のために第1のサブネットワークおよび第2のサブネットワークによって作り出される基準状態分類を割り振る。補足データ割振器はさらに、バリアントアミノ酸配列のために第1のサブネットワークおよび第2のサブネットワークによって作り出されるバリアント状態分類を割り振る。補足データ割振器は、基準アミノ酸配列とアラインメントされた、霊長類PFM、哺乳類PFM、および脊椎動物PFMを割り振る。バリアントアミノ酸配列、割り振られた基準アミノ酸配列、割り振られた基準状態分類およびバリアント状態分類、ならびに割り振られたPFMを処理することに基づいて、バリアントアミノ酸配列を良性または病原性として分類するように訓練される深層畳み込みニューラルネットワークを、多数のプロセッサ上で実行すること。出力プロセッサを通じてバリアントアミノ酸配列に対する病原性スコアを少なくとも報告すること。

【０６６１】

第1のシステム実装形態に対するこの特定の実装セクションにおいて論じられる特徴の各々はこの方法の実装形態に等しく適用される。上で示されたように、すべてのシステム特徴はここで繰り返されず、参照によって繰り返されるものと見なされるべきである。

【０６６２】

他の実装形態は、上で説明された方法を実行するようにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、メモリと、上で説明された方法を実行するためにメモリに記憶された命令を実行するように動作可能な1つまたは複数のプロセッサとを含む、システムを含み得る。

【０６６３】

開示される技術の第2の方法の実装形態は、深層畳み込みニューラルネットワークベースのバリアント病原性分類器を、メモリに結合された多数のプロセッサ上で実行するステップを含む。霊長類PFMおよび哺乳類PFMを生成するために霊長類と哺乳類の2つの配列グループに適用される、多数のプロセッサのうちの少なくとも1つで位置特定的頻度行列(PFMと省略される)生成器を実行すること。入力プロセッサにおいて、標的バリアントアミノ酸の上流および下流の側に各方向への少なくとも25個のアミノ酸があるバリアントアミノ酸配列を受け入れること。一塩基バリアントは標的バリアントアミノ酸を作り出す。バリアントアミノ酸配列とアラインメントされる、標的基準アミノ酸の上流および下流の側に各方向への少なくとも25個のアミノ酸がある基準アミノ酸配列を割り振り、基準アミノ酸配列とアラインメントされる霊長類PFMおよび哺乳類PFMを割り振る、多数のプロセッサのうちの少なくとも1つで補足データ割振器を実行すること。バリアントアミノ酸配列、割

り振られた基準アミノ酸配列、および割り振られたPFMを処理することに基づいて、バリアントアミノ酸配列を良性または病原性として分類するように訓練される深層畳み込みニューラルネットワークを、多数のプロセッサ上で実行すること。出力プロセッサにおいてバリアントアミノ酸配列に対する病原性スコアを少なくとも報告すること。

【０６６４】

第2のシステム実装形態に対するこの特定の実装セクションにおいて論じられる特徴の各々は、この方法の実装形態に等しく適用される。上で示されたように、すべてのシステム特徴はここで繰り返されず、参照によって繰り返されるものと見なされるべきである。

【０６６５】

他の実装形態は、上で説明された方法を実行するようにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、メモリと、上で説明された方法を実行するためにメモリに記憶された命令を実行するように動作可能な1つまたは複数のプロセッサとを含む、システムを含み得る。

【０６６６】

開示される技術のさらに別のシステム実装形態は、一塩基多型(SNPと省略される)病原性分類器を訓練するための大規模な病原性訓練データを生成するシステムを含む。

【０６６７】

図19に示されるように、システムは、良性SNPの訓練セットと、組合せで生成されたSNPの合成セットからカリング(cull)される予測されるエリート病原性SNPの訓練セットとを使用して、メモリに結合された多数のプロセッサ上で実行されるSNP病原性分類器を訓練する。本出願の文脈では、予測されるエリート病原性SNPは、アンサンブルによって出力されるような、平均病原性スコアまたは最大病原性スコアに基づいて各サイクルの終わりに作成/選択されるSNPである。「エリート」という用語は、遺伝的アルゴリズムの語彙から借用され、遺伝的アルゴリズムの出版物において通常与えられる意味を有することが意図される。

【０６６８】

図37、図38、図39、図40、図41、および図42に示されるように、システムはサイクルの中で反復的にエリートセットを構築する。このとき、予測されるSNPがない状態から始めて、合成セットから異常値SNPをカリングすることによって予測されるSNPの完全なセットを累積する。合成セットは、良性セットに存在しない、組合せで生成されるSNPである疑似病原性SNPを備え、異常値SNPがエリートセットへの包含のために合成セットから反復的にカリングされるにつれてセットのメンバー数が減少する。本出願の文脈では、「カリングする」という用語は、以前の集団をフィルタリングすること、新しい集団で置き換えること、更新すること、または選択することを意味する。「カリングする」という用語は、遺伝的アルゴリズムの語彙から借用され、遺伝的アルゴリズムの出版物において通常与えられる意味を有することが意図される。

【０６６９】

図37、図38、図39、図40、図41、および図42に示されるように、システムは、サイクルにおいて反復的に、合成セットから異常値SNPをカリングするために、SNP病原性分類器のアンサンブルを訓練して適用する。これは、良性SNPの一般訓練セット、予測されるエリート病原性SNPの一般訓練セット、および置換を伴わずに合成セットからサンプリングされる疑似病原性SNPの別個の訓練セットを使用して、アンサンブルを訓練することを含む。これはまた、現在のサイクルにおいてアンサンブルを訓練するために使用されなかった合成セットからの少なくともいくつかのSNPをスコアリングするために、訓練されたアンサンブルを適用し、スコアリングされたSNPから、一般エリートセットにおいて累積すべき現在のサイクルの異常値SNPをスコアリングされたSNPから選択するためにスコアを使用することによって、合成セットから異常値SNPをカリングしてカリングされた異常値SNPを一般エリートセットにおいて累積するように、訓練されたアンサンブルを適用することを含む。

【０６７０】

本出願の文脈では、「疑似病原性SNP」は、訓練の目的で病原性としてラベリングされ、訓練の間に置換を伴わずに合成的に生成されたバリアントからサンプリングされるSNPである。

【０６７１】

また、予測されるエリート病原性SNPの訓練セットは、複数のサイクルにわたって反復的に構築される。

【０６７２】

図37、図38、図39、図40、図41、および図42に示されるように、システムは次いで、訓練によって導かれた分類器パラメータ、複数のサイクルにわたって完成され一般良性セットの所定の範囲内にある一般エリートセットと、SNP病原性分類器を訓練するための一般良性セットとを、メモリに記憶する。

【０６７３】

図37、図38、図39、図40、図41、および図42に示されるように、予測されるエリート病原性SNPは、アンサンブルによって予測されるSNPの上位5%である。いくつかの実装形態では、それらは20000個などの固定された数の上位にスコアリングされるSNPである。

【０６７４】

SNP病原性分類器およびSNP病原性分類器のアンサンブルは各々、深層畳み込みニューラルネットワーク(DCNNと省略される)である。アンサンブルは4個から16個のDCNNを含む。図37、図38、図39、図40、図41、および図42に示されるように、アンサンブルは8個のDCNNを含む。

【０６７５】

図37、図38、図39、図40、図41、および図42に示されるように、システムは、サイクルの間のエポックにおいてDCNNのアンサンブルを訓練し、妥当性確認サンプルに対する予測が良性予測と病原性予測の別々の確率分布クラスタを形成するとき、特定のサイクルに対する訓練を終了する。

【０６７６】

図37、図38、図39、図40、図41、および図42に示されるように、システムは、スコアを使用して、DCNNのアンサンブルからのスコアを合計することによって現在のサイクルの異常値SNPを選択する。

【０６７７】

図37、図38、図39、図40、図41、および図42に示されるように、システムは、スコアを使用して、DCNNのアンサンブルによってスコアリングされるSNPの各々に対する最大平均値をとることによって現在のサイクルの異常値SNPを選択する。

【０６７８】

図37、図38、図39、図40、図41、および図42に示されるように、現在のサイクルの間の置換を伴わないサンプリングは、現在のサイクルの間の疑似病原性SNPの互いに素の別々の訓練セットをもたらす。

【０６７９】

システムは、終了条件に達するまでサイクルを続ける。終了条件はサイクルの所定の数であり得る。図37、図38、図39、図40、図41、および図42に示されるように、サイクルの所定の数は21である。

【０６８０】

図37、図38、図39、図40、図41、および図42に示されるように、終了条件は、予測されるエリート病原性セットサイズが良性セットサイズの所定の範囲内にあるときである。

【０６８１】

分類器パラメータは、少なくとも畳み込みフィルタ重みおよび学習率であり得る。

【０６８２】

システムは、SNP病原性分類器としてアンサンブルの中のSNP病原性分類器のうちの1つを選択することができる。選択されるSNP病原性分類器は、最終のサイクルにおいて評価される妥当性確認サンプルについてアンサンブルの中の他のSNP病原性分類器より予測が

優れていた分類器であり得る。

【０６８３】

図37、図38、図39、図40、図41、および図42に示されるように、複数のサイクルにわたって完成する一般エリートセットは、少なくとも400000個の予測されるエリート病原性SNPを有し得る。

【０６８４】

図37、図38、図39、図40、図41、および図42に示されるように、システムは、各サイクルにおいて、予測されるエリート病原性SNPにおける変異率バイアスを防ぐために、良性SNPとサンプリングされた疑似病原性SNPとの間でトリヌクレオチドコンテクストを照合することができる。

【０６８５】

図37、図38、図39、図40、図41、および図42に示されるように、同期セットからの疑似病原性SNPのサンプリングは、各々の連続するサイクルにおいて5%ずつ減少し得る。

【０６８６】

図37、図38、図39、図40、図41、および図42に示されるように、システムは、訓練のために現在のサイクルにおいてサンプリングされる疑似病原性SNPによって現在のサイクルにおいてスコアリングされる合成SNP、予測されるエリート病原性SNP、および訓練のために現在のサイクルにおいて使用される良性SNPをフィルタリングすることができる。

【０６８７】

第1のシステム実装形態に対するこの特定の実装セクションにおいて論じられる特徴の各々が、このシステム実装形態に等しく適用される。上で示されるように、すべてのシステム特徴はここで繰り返されず、参照によって繰り返されるものと見なされるべきである。

【０６８８】

他の実装形態は、上で説明されたシステムの活動を実行するようにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、メモリと、上で説明されたシステムの活動を実行するようにメモリに記憶されている命令を実行するように動作可能な1つまたは複数のプロセッサとを含む、システムを含み得る。

【０６８９】

開示される技術の別の実装形態は、図36に示されるように、畳み込みニューラルネットワーク(CNNと省略される)ベースの半教師あり学習器を含む。

【０６９０】

図36に示されるように、半教師あり学習器は、良性訓練セットおよび病原性訓練セットについて反復的に訓練される、メモリに結合された多数のプロセッサ上で実行されるCNNのアンサンブルを含み得る。

【０６９１】

図36に示されるように、半教師あり学習器は、訓練されたアンサンブルによる合成セットの評価に基づいて病原性訓練セットのセットサイズを次第に増強する、プロセッサのうちの少なくとも1つで実行されるセット増強器を含み得る。

【０６９２】

各反復において、評価は、セット増強器によって病原性訓練セットに追加される、予測されるエリート病原性セットを作り出す。

【０６９３】

半教師あり学習器は、CNN、増強された病原性訓練セット、および良性訓練セットのうちの少なくとも1つを使用して、一塩基多型(SNPと省略される)病原性分類器を構築して訓練する、ビルダを含み得る。

【０６９４】

第1のシステム実装形態に対するこの特定の実装セクションにおいて論じられる特徴の各々が、このシステム実装形態に等しく適用される。上で示されるように、すべてのシス

テム特徴はここで繰り返されず、参照によって繰り返されるものと見なされるべきである。

【０６９５】

他の実装形態は、上で説明されたシステムの活動を実行するようにプロセッサによって実行可能な命令を記憶する非一時的コンピュータ可読記憶媒体を含み得る。さらに別の実装形態は、メモリと、上で説明されたシステムの活動を実行するようにメモリに記憶されている命令を実行するように動作可能な1つまたは複数のプロセッサとを含む、システムを含み得る。

【０６９６】

先行する説明は、開示される技術の作成および使用を可能にするために提示される。開示される実装形態に対する様々な修正が明らかであり、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および範囲から逸脱することなく、他の実装形態および適用例に適用され得る。したがって、開示される技術は、示される実装形態に限定されることは意図されず、本明細書で開示される原理および特徴と一致する最も広い範囲を認められるべきである。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定義される。

【０６９７】

［コンピュータシステム］

図66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の簡略化されたブロック図である。コンピュータシステムは通常、バスサブシステムを介していくつかの周辺デバイスと通信する少なくとも1つのプロセッサを含む。これらの周辺デバイスは、たとえば、メモリデバイスおよびファイルストレージサブシステム、ユーザインターフェース入力デバイス、ユーザインターフェース出力デバイス、ならびにネットワークインターフェースサブシステムを含む、ストレージサブシステムを含み得る。入力デバイスおよび出力デバイスはコンピュータシステムとのユーザの対話を可能にする。ネットワークインターフェースサブシステムは、他のコンピュータシステムにおける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。

【０６９８】

一実装形態では、良性データセット生成器、バリアント病原性分類器、二次構造分類器、溶媒接触性分類器、および半教師あり学習器などのニューラルネットワークは、ストレージサブシステムおよびユーザインターフェース入力デバイスへ通信可能につながれる。

【０６９９】

ユーザインターフェース入力デバイスは、キーボードと、マウス、トラックボール、タッチパッド、またはグラフィクスタブレットなどのポインティングデバイスと、ディスプレイに組み込まれたタッチスクリーンと、音声認識システムおよびマイクロフォンなどのオーディオ入力デバイスと、他のタイプの入力デバイスとを含み得る。一般に、「入力デバイス」という用語の使用は、コンピュータシステムへ情報を入力するためのすべての可能なタイプのデバイスおよび方式を含むことが意図される。

【０７００】

ユーザインターフェース出力デバイスは、ディスプレイサブシステム、プリンタ、faxマシン、またはオーディオ出力デバイスなどの非視覚的ディスプレイを含み得る。ディスプレイサブシステムは、陰極線管(CRT)、液晶ディスプレイ(LCD)などのフラットパネルデバイス、プロジェクションデバイス、または可視の画像を創造するための何らかの他の機構を含み得る。ディスプレイサブシステムはまた、オーディオ出力デバイスなどの非視覚ディスプレイを提供することができる。一般に、「出力デバイス」という用語の使用は、コンピュータシステムから情報をユーザまたは別の機械もしくはコンピュータシステムに出力するためのすべての可能なタイプのデバイスおよび方式を含むことが意図される。

【０７０１】

ストレージサブシステムは、本明細書で説明されるモジュールおよび方法の一部またはすべての機能を提供する、プログラミングおよ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、プロセッサだけによって、または他のプロセッサと組み合わせて実行される。

【０７０２】

ストレージサブシステムにおいて使用されるメモリは、プログラム実行の間の命令およびデータの記憶のためのメインランダムアクセスメモリ(RAM)と、固定された命令が記憶される読取り専用メモリ(ROM)とを含む、いくつかのメモリを含み得る。ファイルストレージサブシステムは、プログラムおよびデータファイルのための永続的なストレージ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メディアを伴うフロッピーディスクドライブ、CD-ROMドライブ、光学ドライブ、または取り外し可能なメディアカートリッジを含み得る。いくつかの実装形態の機能を実装するモジュールは、ストレージサブシステムの中の、または他のプロセッサによってアクセス可能な他の機械の中の、ファイルストレージサブシステムによって記憶され得る。

【０７０３】

バスサブシステムは、コンピュータシステムの様々な構成要素およびサブシステムに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、バスサブシステムの代替的な実装形態は複数のバスを使用することができる。

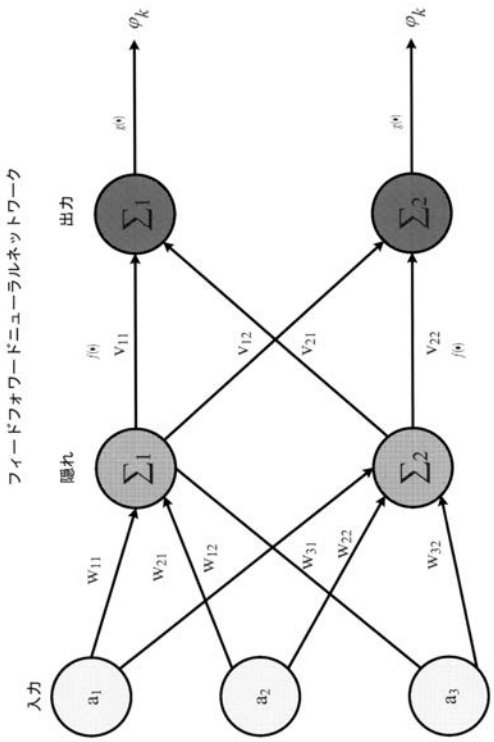
【０７０４】

コンピュータシステム自体が、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、テレビジョン、メインフレーム、サーバファーム、緩やかにネットワーク化されたコンピュータの広く分布するセット、または、任意の他のデータ処理システムもしくはユーザデバイスを含む、様々なタイプであってよい。コンピュータおよびネットワークの変わり続ける性質により、図66に示されるコンピュータシステムの記述は、開示される技術を例示することを目的とする特定の例としてのみ意図されている。図66に示されるコンピュータシステムより多数または少数の構成要素を有する、コンピュータシステムの多くの他の構成が可能である。

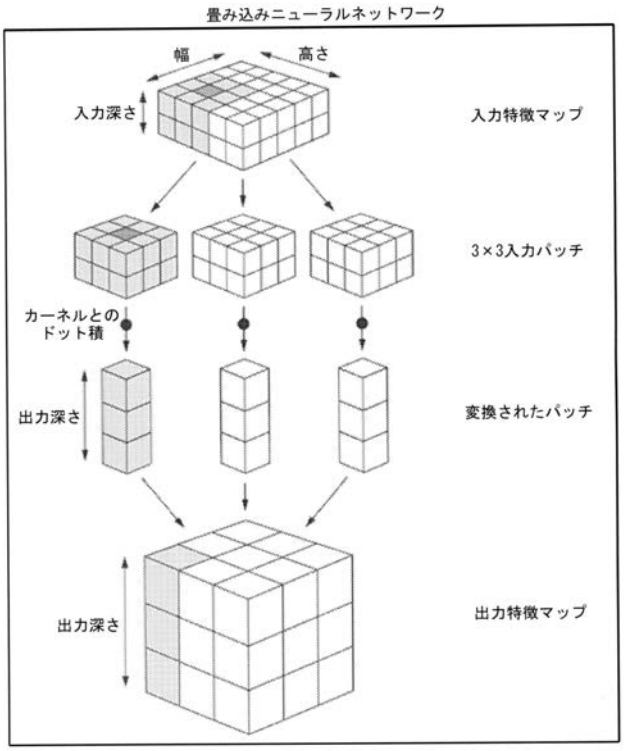
【０７０５】

深層学習プロセッサは、GPUまたはFPGAであってよく、Google Cloud Platform、Xilinx、およびCirrascale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てよい。深層学習プロセッサの例には、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のStratix V FPGA、GraphcoreのIntelligent Processor Unit(IPU)、Snapdragonプロセッサを用いたQualcommのZerothプラットフォーム、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などがある。

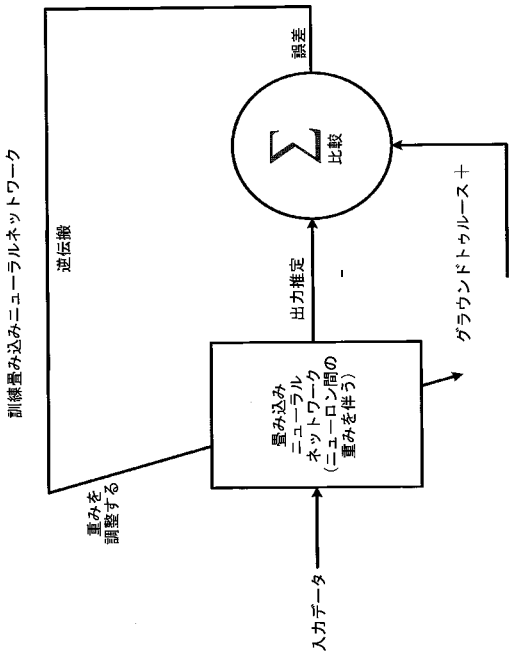
【図１Ａ】



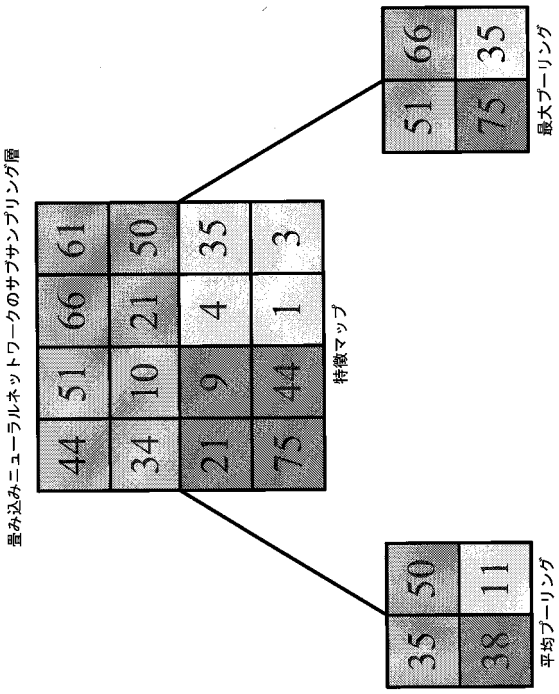
【図１Ｂ】



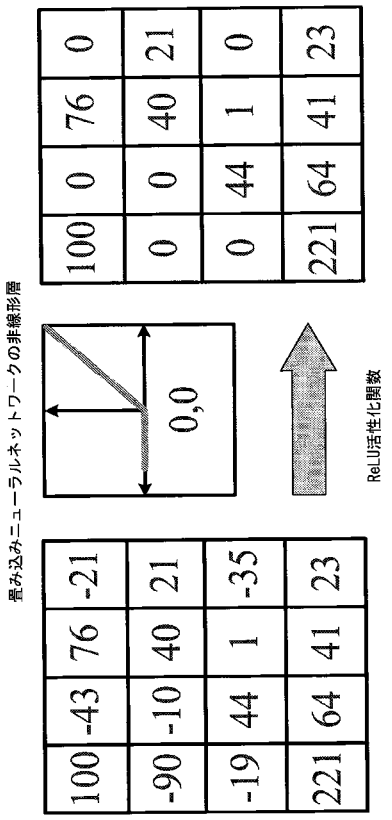
【図１Ｃ】



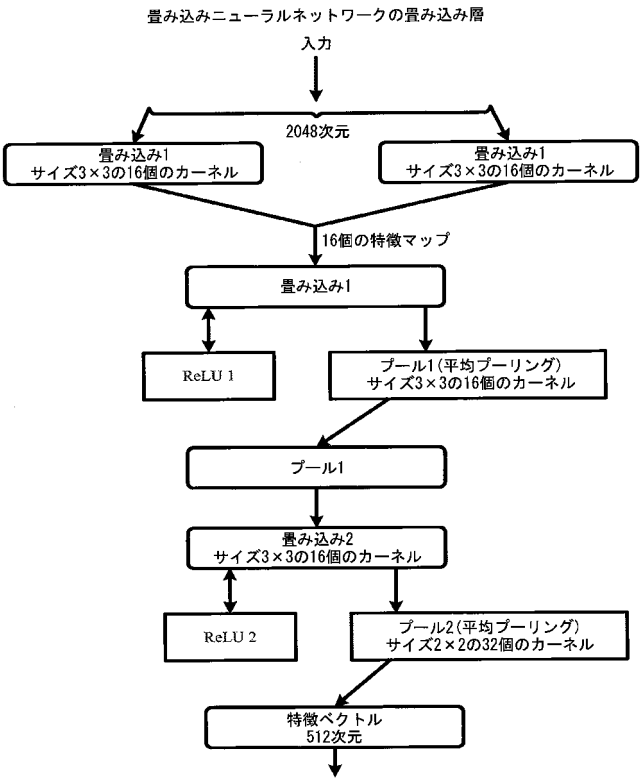
【図１Ｄ】



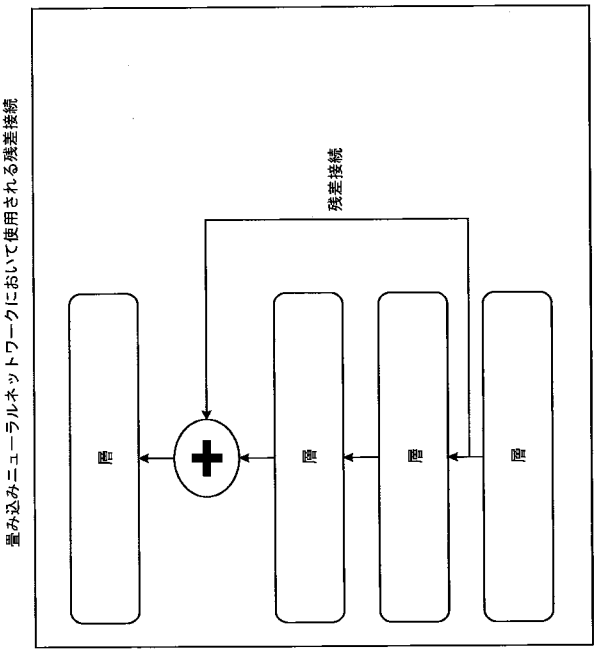
【図１Ｅ】



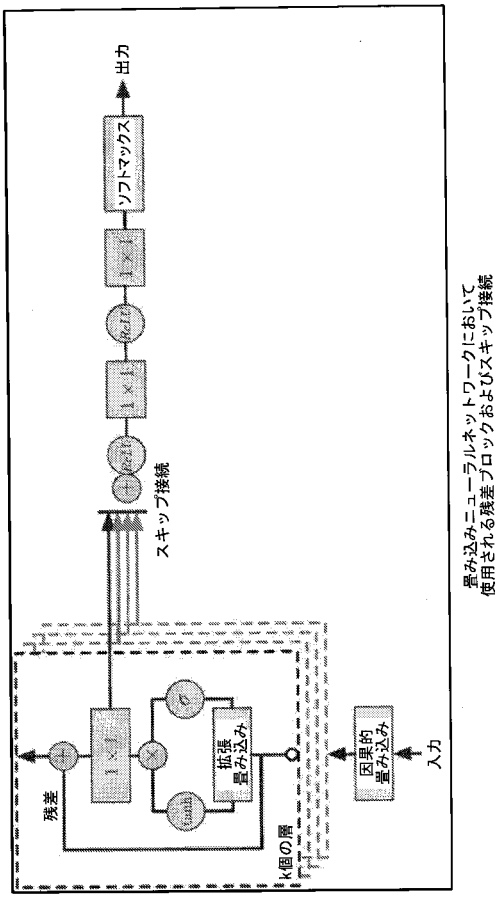
【図１Ｆ】



【図１Ｇ】



【図１Ｈ】



【図１Ｉ】

畳み込みニューラルネットワークを用いたバッチ正規化フォワードパス

$$\mu_{\mathcal{B}} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(\ell-1)}$$

$$\sigma_{\mathcal{B}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^{(\ell-1)} - \mu_{\mathcal{B}})^2$$

$$\hat{x}^{(\ell-1)} = \frac{x^{(\ell-1)} - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

$$x^{(\ell)} = \gamma^{(\ell)}\hat{x}^{(\ell-1)} + \beta^{(\ell)}$$

【図１Ｊ】

バッチ正規化-畳み込みニューラルネットワークを用いた推論

$$\hat{x}^{(\ell-1)} = \frac{x^{(\ell-1)} - \mu_{\mathcal{D}}}{\sqrt{\sigma_{\mathcal{D}}^2 + \epsilon}}$$

$$x_i^{(\ell)} = \gamma^{(\ell)}\hat{x}_i^{(\ell-1)} + \beta^{(\ell)}$$

【図１Ｋ】

畳み込みニューラルネットワークを用いたバッチ正規化バックワードパス

$$\nabla_{\gamma^{(\ell)}}\mathcal{L} = \sum_{i=1}^{n}\left(\nabla_{x^{(\ell+1)}}\mathcal{L}\right)_i \cdot \hat{x}_i^{(\ell)}$$

$$\nabla_{\beta^{(\ell)}}\mathcal{L} = \sum_{i=1}^{n}\left(\nabla_{x^{(\ell+1)}}\mathcal{L}\right)_i$$

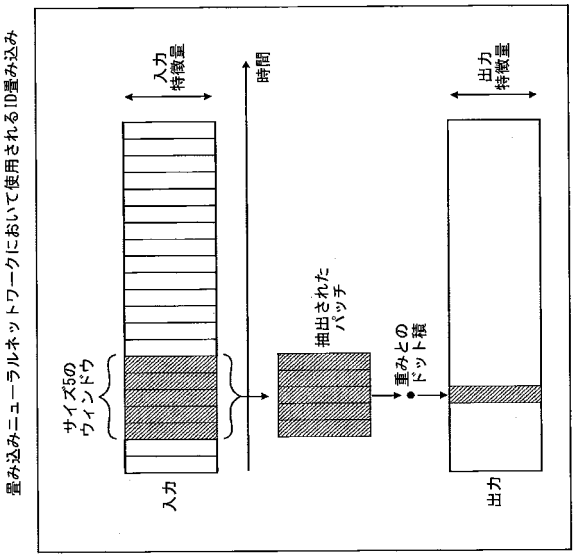
【図１Ｌ】

畳み込み層におけるバッチ正規化

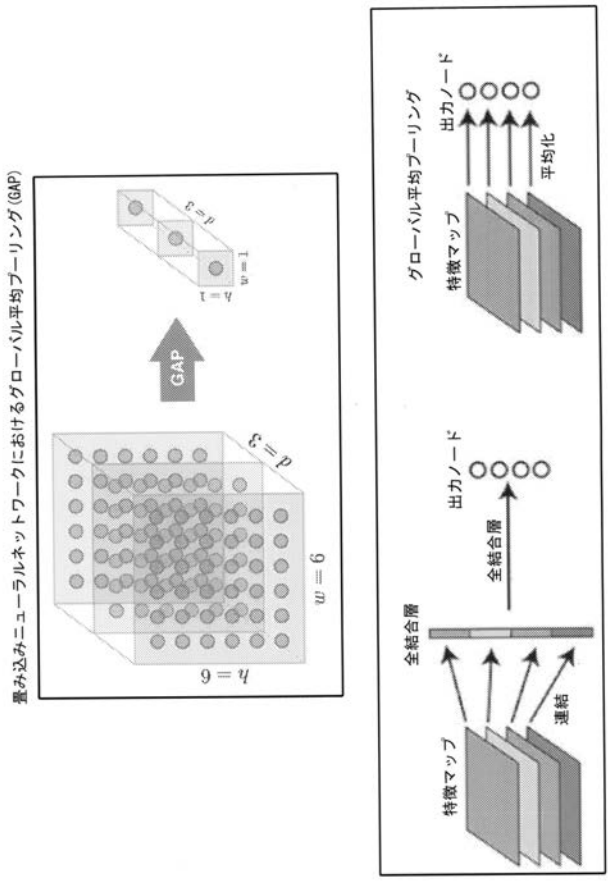
```
conv_model.add(layers.Conv2D(32, 3, activation='relu'))  ←―― 畳み込み層の後
conv_model.add(layers.BatchNormalization())

dense_model.add(layers.Dense(32, activation='relu'))  ←―― 密層の後
dense_model.add(layers.BatchNormalization())
```

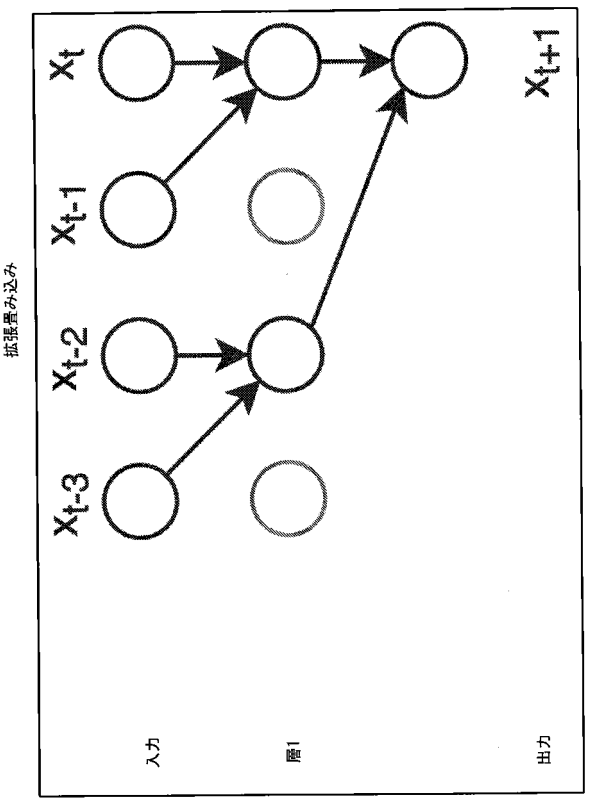
【図１Ｍ】



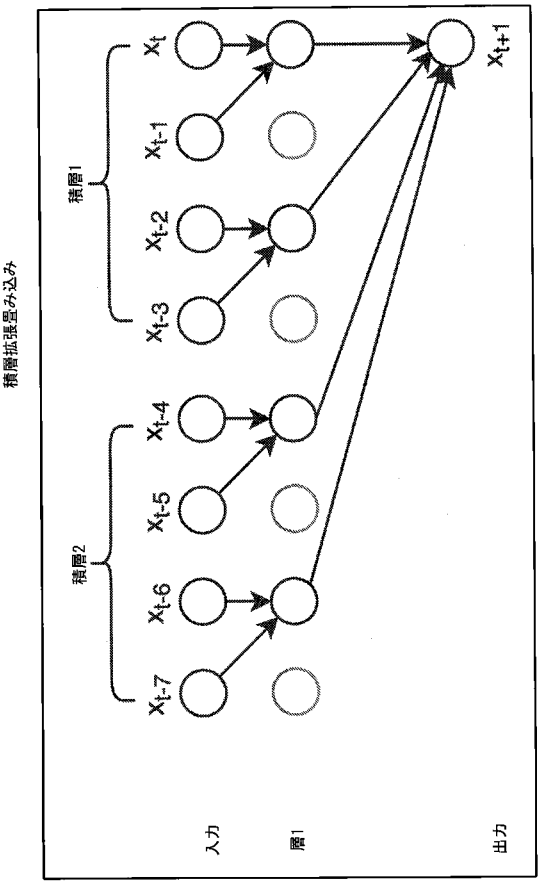
【図１Ｎ】



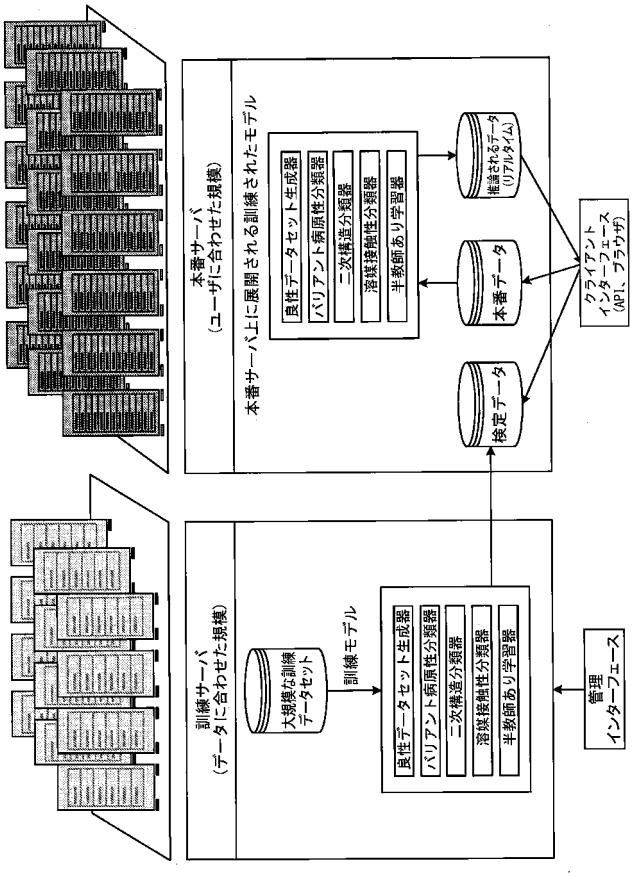
【図１Ｏ】



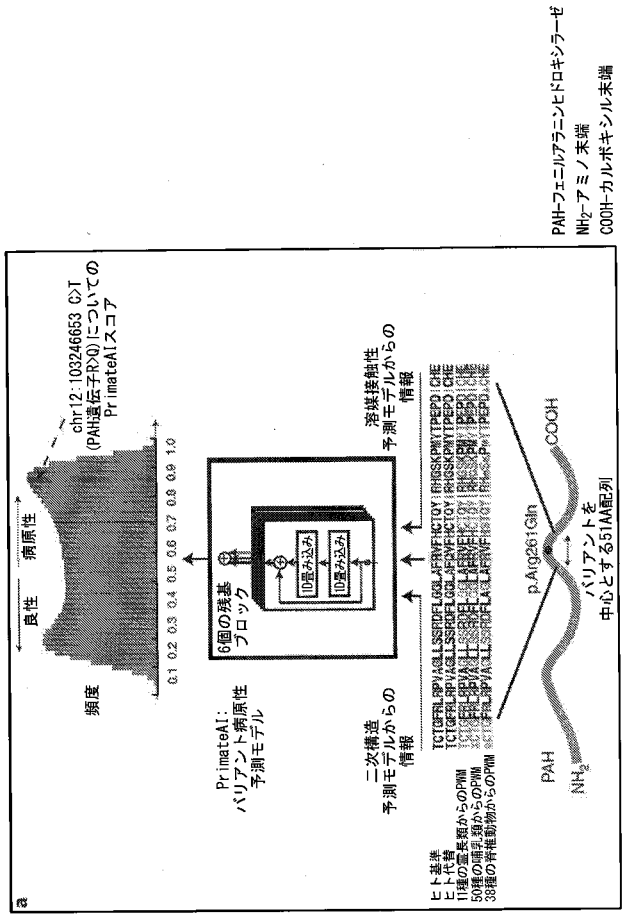
【図１Ｐ】



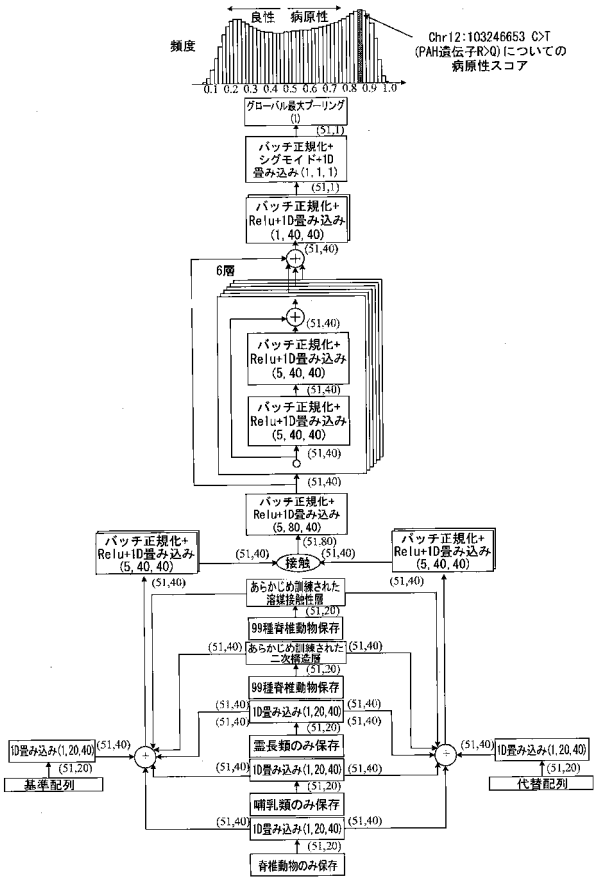
【図１Ｑ】



【図２】



【図３】



【図４Ａ】

層	層への入力	タイプ	カーネルの数&ウィンドウサイズ	形状	活性化
層1a	基準配列	畳み込み1D	40,1	(51,40)	線形
層1b	代替配列	畳み込み1D	40,1	(51,40)	線形
層1c	霊長類保存	畳み込み1D	40,1	(51,40)	線形
	哺乳類保存	畳み込み1D	40,1	(51,40)	線形
層1d	脊椎動物保存	畳み込み1D	40,1	(51,40)	線形
層1e あらかじめ訓練された二次構造予測モデル(訓練可能ネットワーク)	霊長類+哺乳類+脊椎動物保存	-	-	(51,40)	-
あらかじめ訓練された溶媒接触性モデル(訓練可能ネットワーク)	霊長類+哺乳類+脊椎動物保存	-	-	(51,40)	-
層1f	あらかじめ訓練された構造予測モデル出力	畳み込み1D	40,1	(51,40)	線形
層1g	あらかじめ訓練された溶媒接触性モデル出力	畳み込み1D	40,1	(51,40)	線形
マージ1a	層1a、層1c、層1d、層1e、層1f、層1g	加算	-	(51,40)	-
マージ1b	層1b、層1c、層1d、層1e、層1f、層1g	加算	-	(51,40)	-
層2a	マージ1a	バッチ正規化	-	(51,40)	-
層3a	層2a	活性化	-	(51,40)	線形
層4a	層3a	畳み込み1D	40,5	(51,40)	
層5a	層4a	バッチ正規化	-	(51,40)	
層6a	層5a	活性化	-	(51,40)	
層7a	層6a	畳み込み1D	40,5	(51,40)	線形
層2b	マージ1b	バッチ正規化	-	(51,40)	-
層3b	層2b	活性化	-	(51,40)	線形
層4b	層3b	畳み込み1D	40,5	(51,40)	

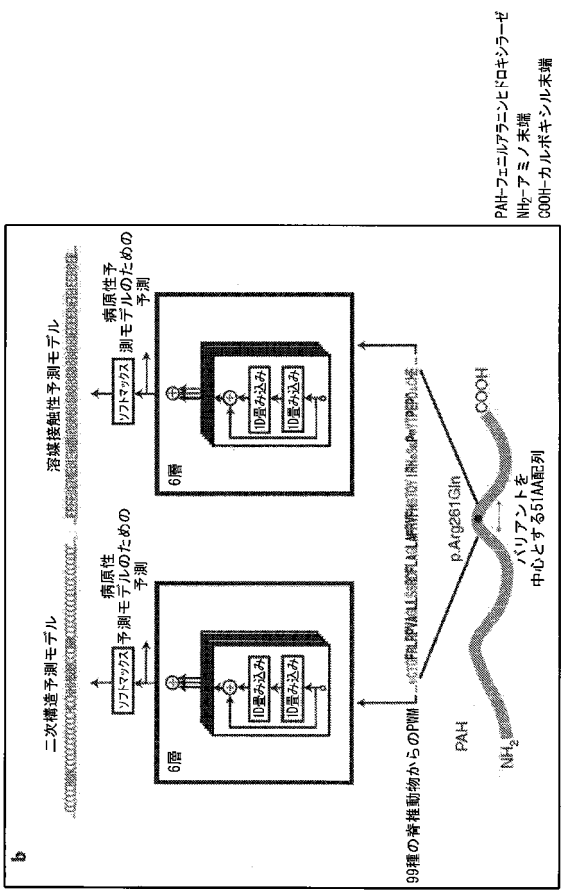
層5b	層4b	バッチ正規化	-	(51,40)	-
層6b	層5b	活性化	-	(51,40)	Relu
層7b	層6b	畳み込み1D	40,5	(51,40)	線形
マージ2a	層7a、層7b	連結	-	(51,80)	-
マージ2b	層7a、層7b	連結	-	(51,80)	-
層8a	マージ2a	畳み込み1D	40,5	(51,40)	線形
層8b	マージ2b	畳み込み1D	40,5	(51,40)	線形
層9	層8a	バッチ正規化	-	(51,40)	-
層10	層9	活性化	-	(51,40)	Relu
層11	層10	畳み込み1D	40,5	(51,40)	線形
層12	層11	バッチ正規化	-	(51,40)	-
層13	層12	活性化	-	(51,40)	Relu
層14	層13	畳み込み1D	40,5	(51,40)	線形
マージ3	層8a、層14	加算	-	(51,40)	-
層15	マージ3	バッチ正規化	-	(51,40)	-
層16	層15	活性化	-	(51,40)	Relu
層17	層16	畳み込み1D	40,5	(51,40)	線形
層18	層17	バッチ正規化	-	(51,40)	-
層19	層18	活性化	-	(51,40)	Relu
層20	層19	畳み込み1D	40,5	(51,40)	線形
マージ4	マージ3、層20	加算	-	(51,40)	-
層21	マージ4	畳み込み1D	40,1	(51,40)	線形
層22	マージ4	バッチ正規化	-	(51,40)	-
層23	層22	活性化	-	(51,40)	Relu
層24	層23	畳み込み1D	40,5	(51,40)	線形
層25	層24	バッチ正規化	-	(51,40)	-
層26	層25	活性化	-	(51,40)	Relu
層27	層26	畳み込み1D	40,5	(51,40)	線形
マージ5	マージ4、層27	加算	-	(51,40)	-
層28	マージ5	バッチ正規化	-	(51,40)	-
層29	層28	活性化	-	(51,40)	Relu
層30	層29	畳み込み1D	40,5	(51,40)	線形
層31	層30	バッチ正規化	-	(51,40)	-
層32	層31	活性化	-	(51,40)	Relu

【図４Ｂ】

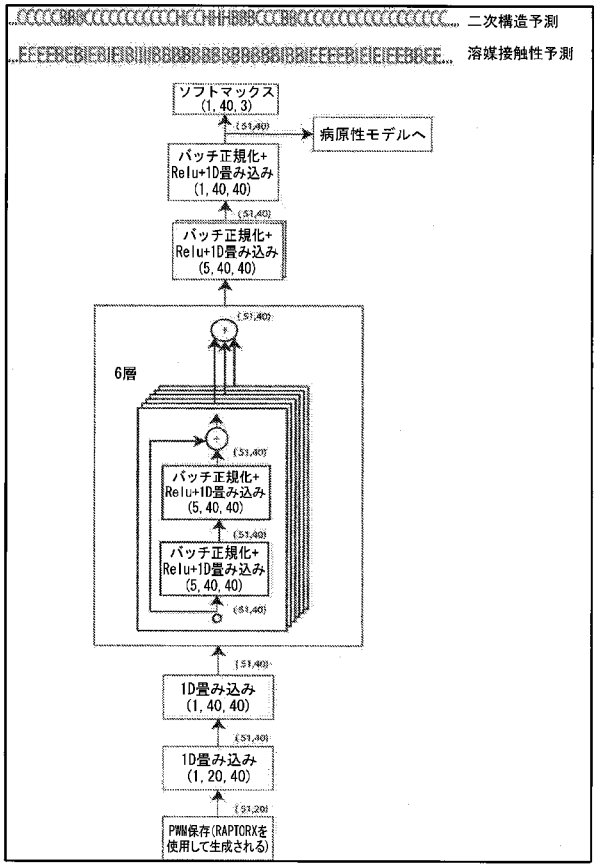
【図４Ｃ】

層33	層32	畳み込み1D	40,5	(51,40)	線形
マージ6	マージ5、層33	加算	-	(51,40)	-
層34	マージ6	畳み込み1D	40,1	(51,40)	線形
層35	マージ6	バッチ正規化	-	(51,40)	-
層36	層35	活性化	-	(51,40)	Relu
層37	層36	畳み込み1D	40,5	(51,40)	線形
層38	層37	バッチ正規化	-	(51,40)	-
層39	層38	活性化	-	(51,40)	Relu
層40	層39	畳み込み1D	40,5	(51,40)	線形
マージ7	マージ6、層40	加算	-	(51,40)	-
層41	マージ7	バッチ正規化	-	(51,40)	-
層42	層41	活性化	-	(51,40)	Relu
層43	層42	畳み込み1D	40,5	(51,40)	線形
層44	層43	バッチ正規化	-	(51,40)	-
層45	層44	活性化	-	(51,40)	Relu
層46	層45	畳み込み1D	40,5	(51,40)	線形
マージ8	マージ7、層46	加算	-	(51,40)	-
層47	マージ8	畳み込み1D	40,1	(51,40)	線形
マージ9	層8b、層21、層34、層47	加算	-	(51,40)	-
層48	マージ10	バッチ正規化	-	(51,40)	-
層49	層48	活性化	-	(51,40)	Relu
層50	層49	畳み込み1D	40,1	(51,40)	線形
層51	層50	バッチ正規化	-	(51,40)	-
層52	層51	活性化	-	(51,40)	Relu
層53	層52	畳み込み1D	40,1	(51,40)	線形
マージ10	マージ9、層53	加算	-	(51,40)	-
層54	マージ10	畳み込み1D	1,1	(51,1)	シグモイド
出力層	層54	グローバル最大プーリング1D			1 -

【図５】



【図６】



【図７Ａ】

層	層への入力	タイプ	カーネルの数&カーネルウィンドウサイズ	形状	活性化
層1a	入力PSFM	畳み込み1D	40,1	(51,40)	線形
層2a	層1a	畳み込み1D	40,1	(51,40)	
層1b	入力PSFM	畳み込み1D	40,1	(51,40)	線形
層2b	層1b	畳み込み1D	40,1	(51,40)	線形
層3	層2a	バッチ正規化	-	(51,40)	-
層4	層3	活性化	-	(51,40)	Relu
層5	層4	畳み込み1D	40,5	(51,40)	線形
層6	層5	バッチ正規化	-	(51,40)	-
層8	層7	活性化	-	(51,40)	Relu
層9	層8	畳み込み1D	40,5	(51,40)	線形
マージ1	層2a、層9	加算	-	(51,40)	-
層10	マージ1	バッチ正規化	-	(51,40)	-
層11	層10	活性化	-	(51,40)	Relu
層12	層11	畳み込み1D	40,5	(51,40)	線形
層13	層12	バッチ正規化	-	(51,40)	-
層14	層13	活性化	-	(51,40)	Relu
層15	層14	畳み込み1D	40,5	(51,40)	線形
マージ2	マージ1、層15	加算	-	(51,40)	-
層16	マージ2	畳み込み1D	40,1	(51,40)	線形
層17	マージ2	バッチ正規化	-	(51,40)	-
層18	層17	活性化	-	(51,40)	Relu
層19	層18	畳み込み1D	40,5	(51,40)	線形
層20	層19	バッチ正規化	-	(51,40)	-
層21	層20	活性化	-	(51,40)	Relu
層22	層21	畳み込み1D	40,5	(51,40)	線形
マージ3	マージ2、層22	加算	-	(51,40)	-

【図７Ｂ】

層23	マージ3	バッチ正規化	-	(51,40)	-
層24	層23	活性化	-	(51,40)	Relu
層25	層24	畳み込み1D	40,5	(51,40)	線形
層26	層25	バッチ正規化	-	(51,40)	-
層27	層26	活性化	-	(51,40)	Relu
層28	層27	畳み込み1D	40,5	(51,40)	線形
マージ4	マージ3、層28	加算	-	(51,40)	-
層29	マージ4	畳み込み1D	40,1	(51,40)	線形
層30	マージ4	バッチ正規化	-	(51,40)	-
層31	層30	活性化	-	(51,40)	Relu
層32	層31	畳み込み1D	40,5	(51,40)	線形
層33	層32	バッチ正規化	-	(51,40)	-
層34	層33	活性化	-	(51,40)	Relu
層35	層34	畳み込み1D	40,5	(51,40)	線形
マージ5	マージ4、層35	加算	-	(51,40)	-
層36	マージ5	バッチ正規化	-	(51,40)	-
層37	層36	活性化	-	(51,40)	Relu
層38	層37	畳み込み1D	40,5	(51,40)	線形
層39	層38	バッチ正規化	-	(51,40)	-
層40	層39	活性化	-	(51,40)	Relu
層41	層40	畳み込み1D	40,5	(51,40)	線形
マージ6	マージ5、層41	加算	-	(51,40)	-
層42	マージ6	畳み込み1D	40,1	(51,40)	線形
マージ7	層2b、層16、層29、層42	加算	-	(51,40)	-
層43	マージ7	バッチ正規化	-	(51,40)	-
層44	層43	活性化	-	(51,40)	Relu
層45	マージ44	畳み込み1D	40,1	(51,40)	線形
層46	層45	バッチ正規化	-	(51,40)	-
層47	層46	活性化	-	(51,40)	Relu
層48	層47	畳み込み1D	40,1	(51,40)	線形
マージ8	マージ7、層48	加算	-	(51,40)	-
出力層	マージ8	畳み込み1D	3,1	(51,3)	ソフトマックス

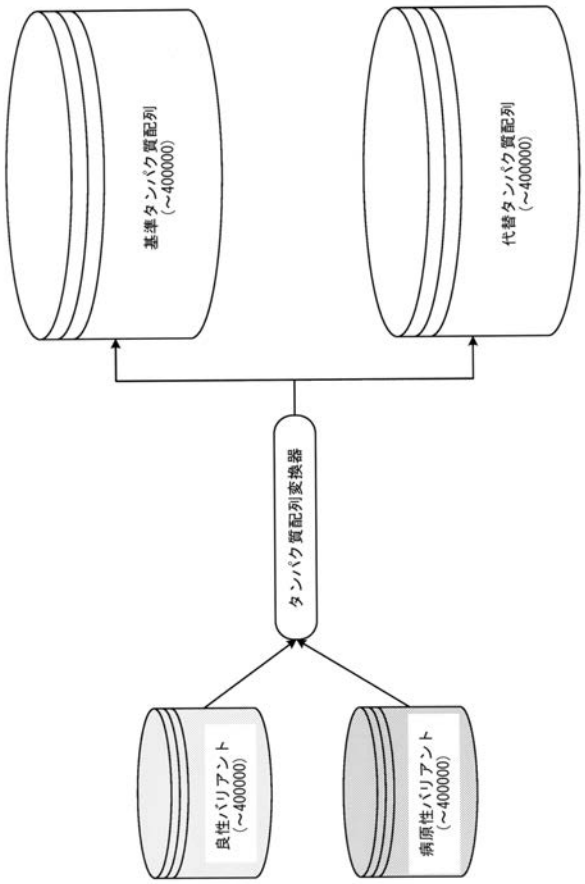
【図８Ａ】

層	層への入力	タイプ	カーネルの数&カーネルウィンドウサイズ	形状	活性化
層1a	入力PSFM	畳み込み1D	40,1	(51,40)	線形
層2a	層1a	畳み込み1D	40,1	(51,40)	線形
層1b	入力PSFM	畳み込み1D	40,1	(51,40)	線形
層2b	層1b	畳み込み1D	40,1	(51,40)	線形
層3	層2a	バッチ正規化	-	(51,40)	-
層4	層3	活性化	-	(51,40)	Relu
層5	層4	畳み込み1D	40,5	(51,40)	線形
層6	層5	バッチ正規化	-	(51,40)	-
層8	層7	活性化	-	(51,40)	Relu
層9	層8	畳み込み1D	40,5	(51,40)	線形
マージ1	層2a、層9	加算	-	(51,40)	-
層10	マージ1	バッチ正規化	-	(51,40)	-
層11	層10	活性化	-	(51,40)	Relu
層12	層11	畳み込み1D	40,5	(51,40)	線形
層13	層12	バッチ正規化	-	(51,40)	-
層14	層13	活性化	-	(51,40)	Relu
層15	層14	畳み込み1D	40,5	(51,40)	線形
マージ2	マージ1、層15	加算	-	(51,40)	-
層16	マージ2	畳み込み1D	40,1	(51,40)	線形
層17	マージ2	バッチ正規化	-	(51,40)	-
層18	層17	活性化	-	(51,40)	Relu
層19	層18	畳み込み1D	40,5	(51,40)	線形
層20	層19	バッチ正規化	-	(51,40)	-
層21	層20	活性化	-	(51,40)	Relu
層22	層21	畳み込み1D	40,5	(51,40)	線形
マージ3	マージ2、層22	加算	-	(51,40)	-

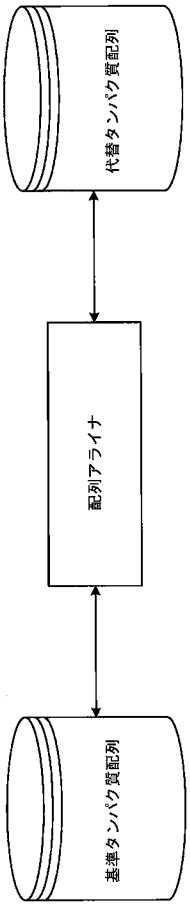
【図８Ｂ】

層23	マージ3	バッチ正規化	-	(51,40)	-
層24	層23	活性化	-	(51,40)	Relu
層25	層24	畳み込み1D	40,5	(51,40)	線形
層26	層25	バッチ正規化	-	(51,40)	-
層27	層26	活性化	-	(51,40)	Relu
層28	層27	畳み込み1D	40,5	(51,40)	線形
マージ4	マージ3、層28	加算	-	(51,40)	-
層29	マージ4	畳み込み1D	40,1	(51,40)	線形
層30	マージ4	バッチ正規化	-	(51,40)	-
層31	層30	活性化	-	(51,40)	Relu
層32	層31	畳み込み1D	40,5	(51,40)	線形
層33	層32	バッチ正規化	-	(51,40)	-
層34	層33	活性化	-	(51,40)	Relu
層35	層34	畳み込み1D	40,5	(51,40)	線形
マージ5	マージ4、層35	加算	-	(51,40)	-
層36	マージ5	バッチ正規化	-	(51,40)	-
層37	層36	活性化	-	(51,40)	Relu
層38	層37	畳み込み1D	40,5	(51,40)	線形
層39	層38	バッチ正規化	-	(51,40)	-
層40	層39	活性化	-	(51,40)	Relu
層41	層40	畳み込み1D	40,5	(51,40)	線形
マージ6	マージ5、層41	加算	-	(51,40)	-
層42	マージ6	畳み込み1D	40,1	(51,40)	線形
マージ7	層2b、層16、層29、層42	加算	-	(51,40)	-
層43	マージ7	バッチ正規化	-	(51,40)	-
層44	層43	活性化	-	(51,40)	Relu
層45	マージ44	畳み込み1D	40,1	(51,40)	線形
層46	層45	バッチ正規化	-	(51,40)	-
層47	層46	活性化	-	(51,40)	Relu
層48	層47	畳み込み1D	40,1	(51,40)	線形
マージ8	マージ7、層48	加算	-	(51,40)	-
出力層	マージ8	畳み込み1D	3,1	(51,3)	ソフトマックス

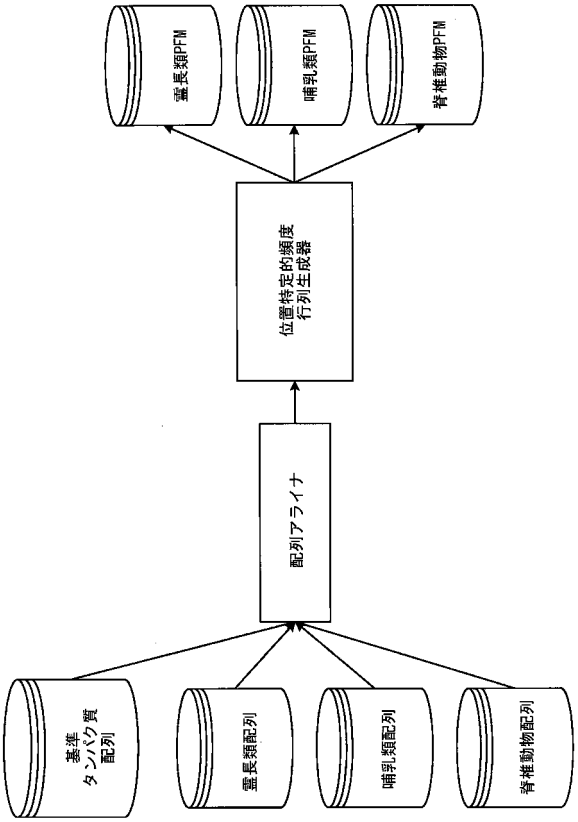
【図９】



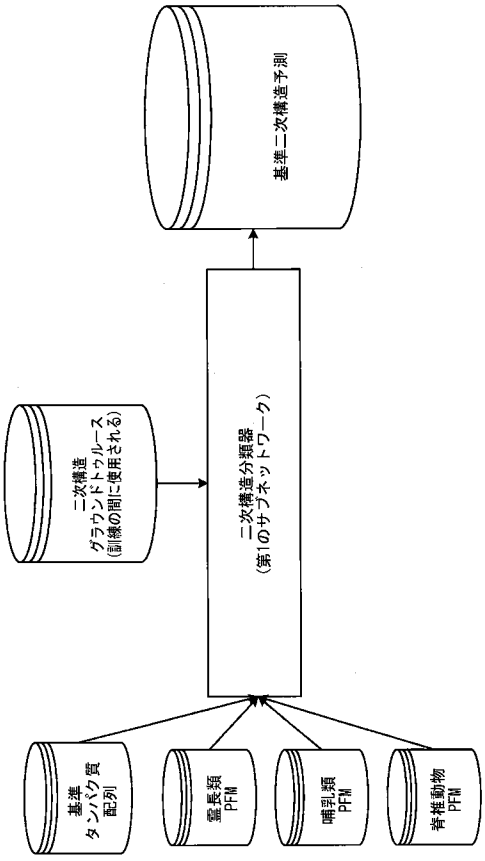
【図１０】



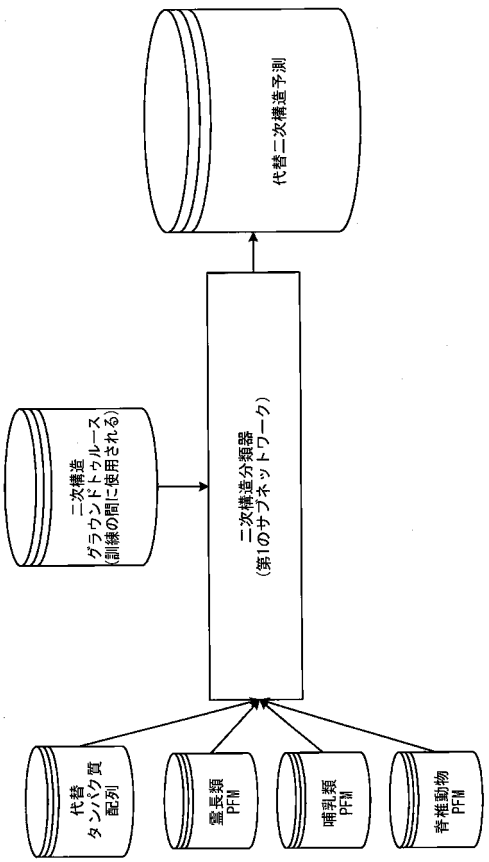
【図１１】



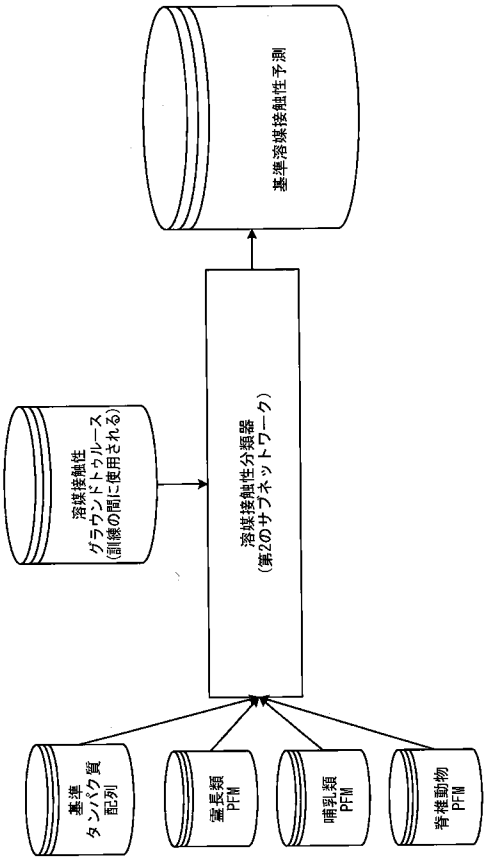
【図１２】



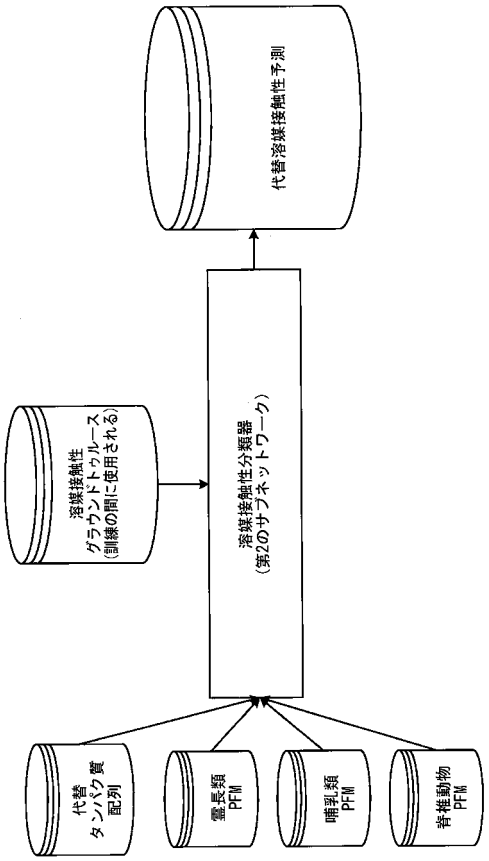
【図１３】



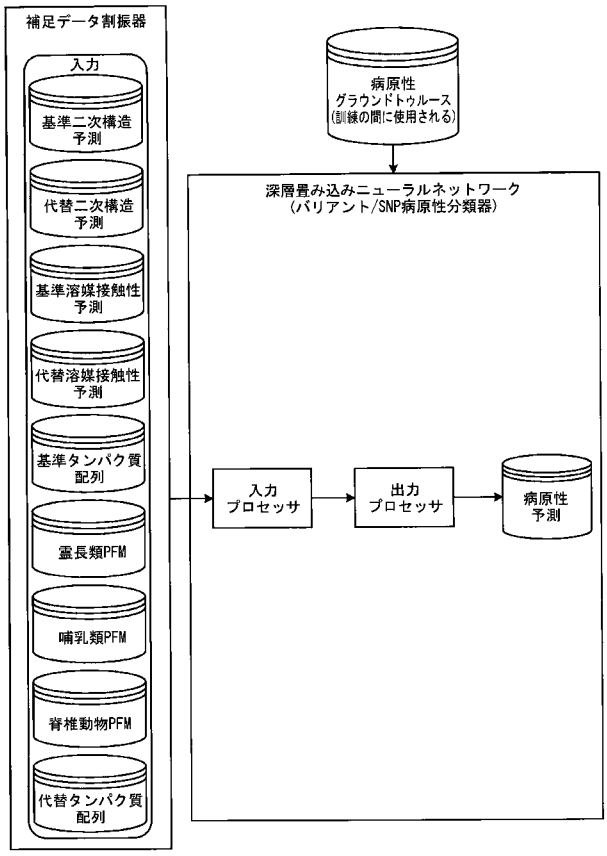
【図１４】



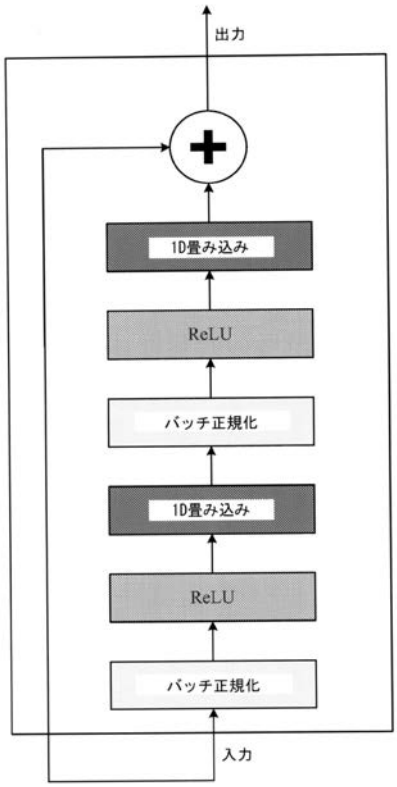
【図１５】



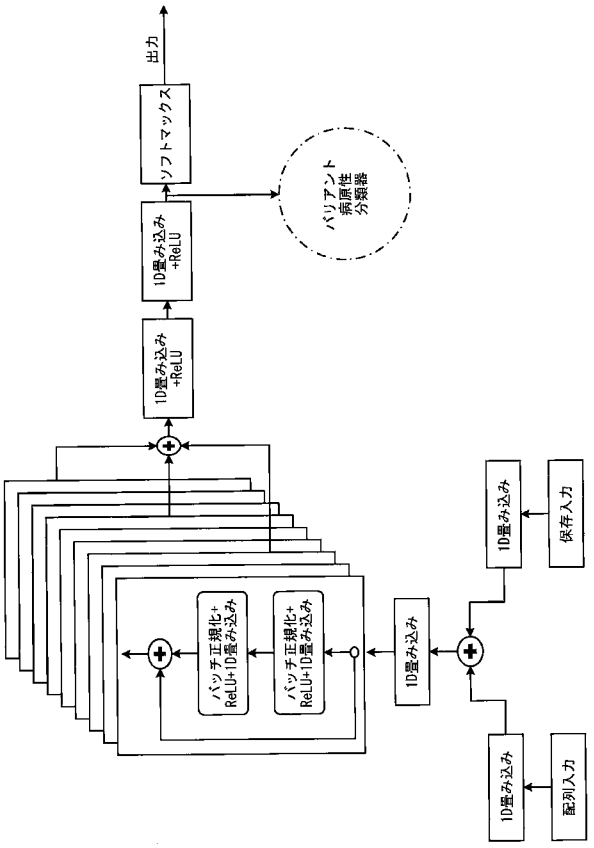
【図１６】



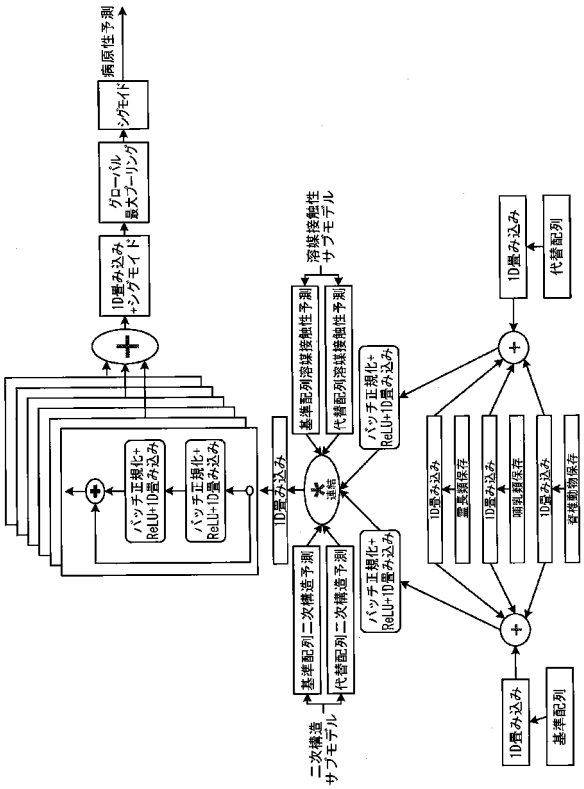
【図１７】



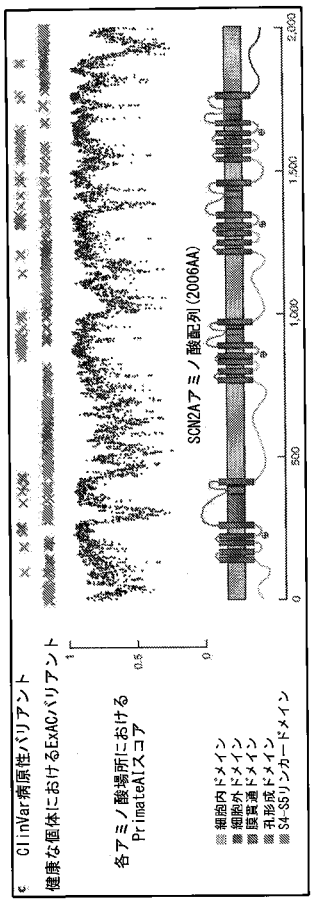
【図１８】



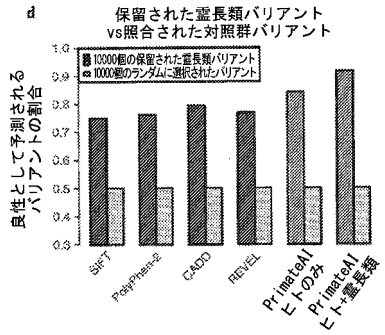
【図１９】



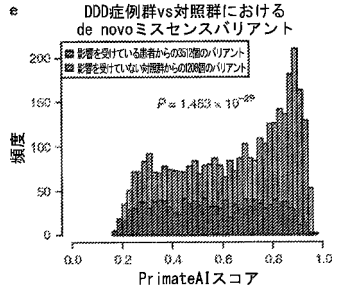
【図２０】



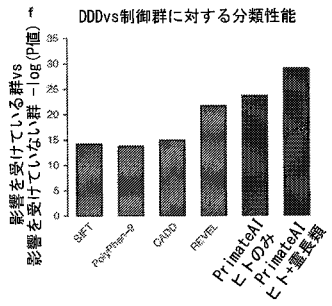
【図２１Ｄ】



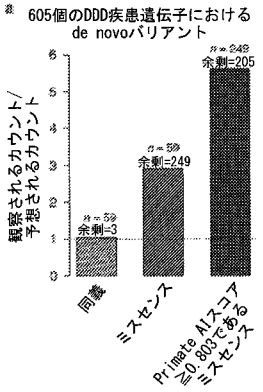
【図２１Ｅ】



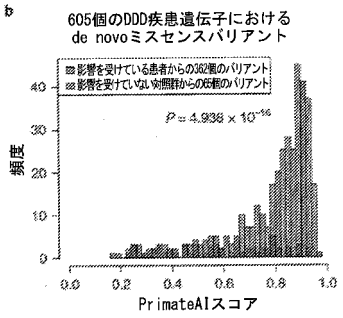
【図２１Ｆ】



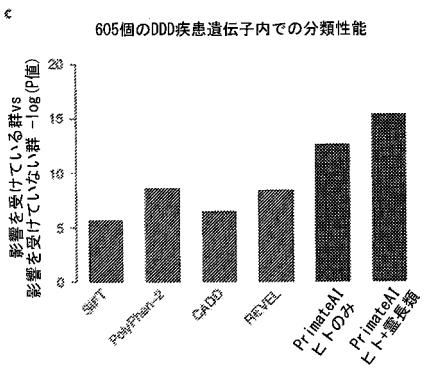
【図２２Ａ】



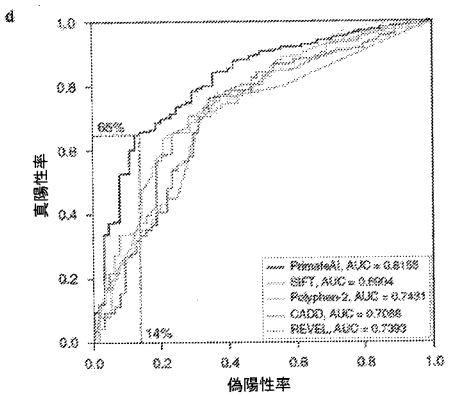
【図２２Ｂ】



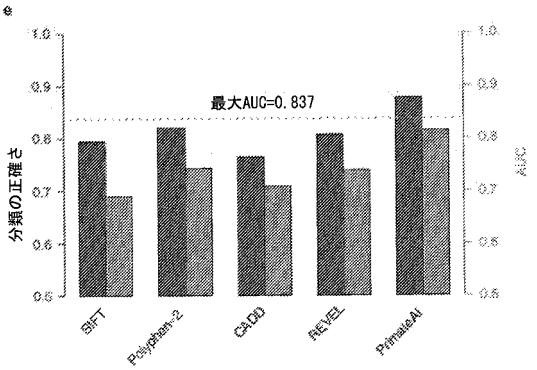
【図２２Ｃ】



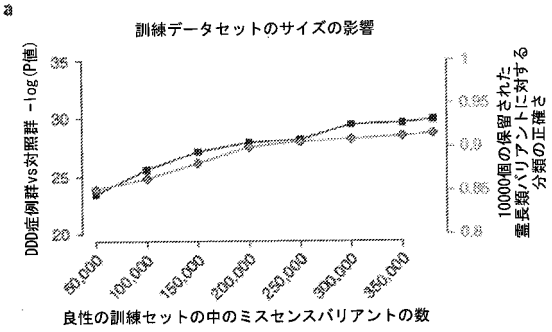
【図２２Ｄ】



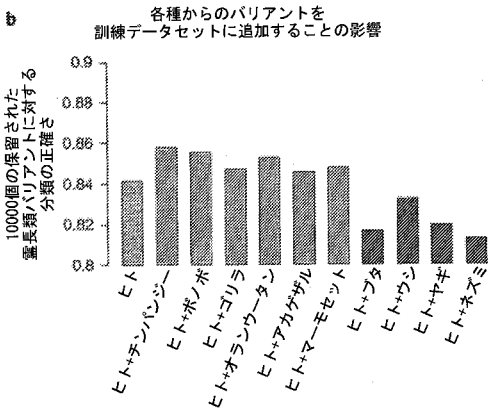
【図２２Ｅ】



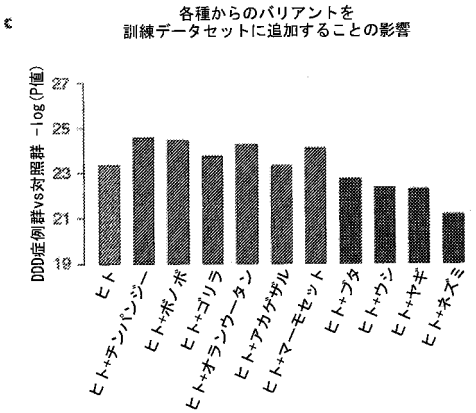
【図２３Ａ】



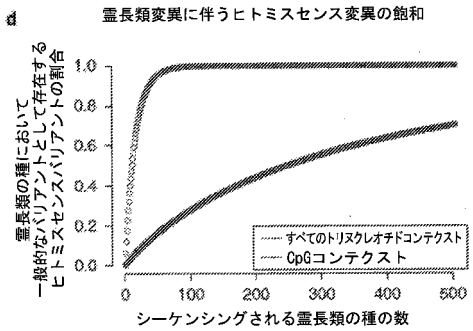
【図２３Ｂ】



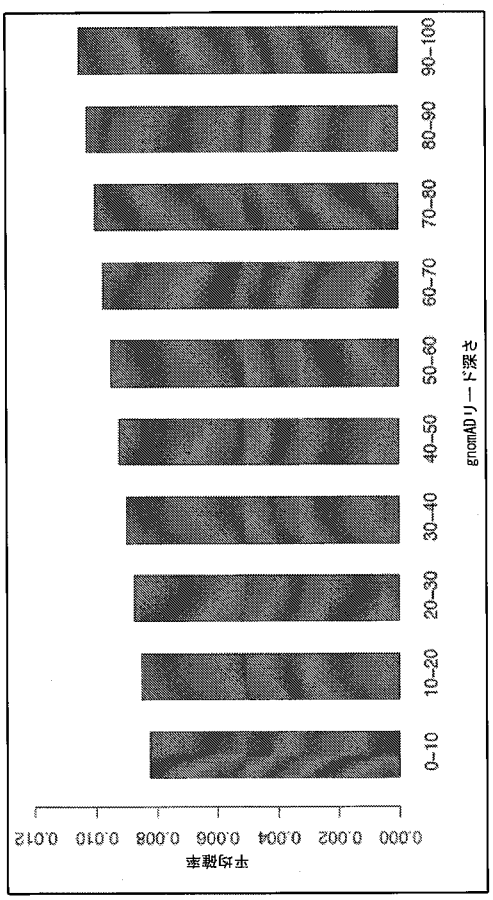
【図２３Ｃ】



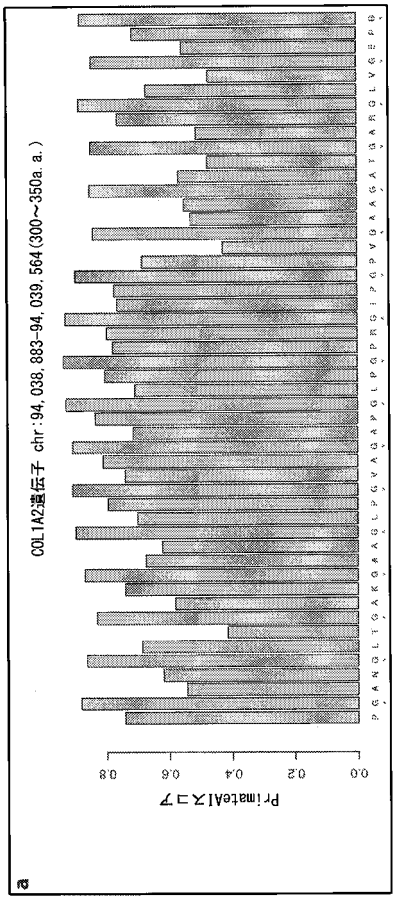
【図２３Ｄ】



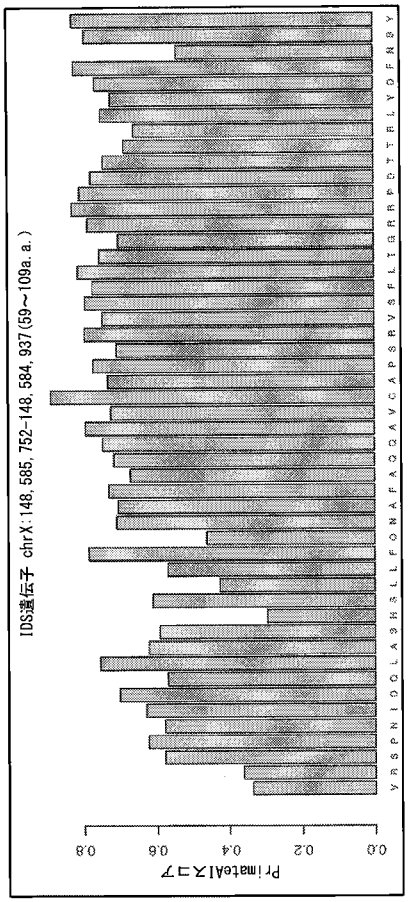
【図２４】



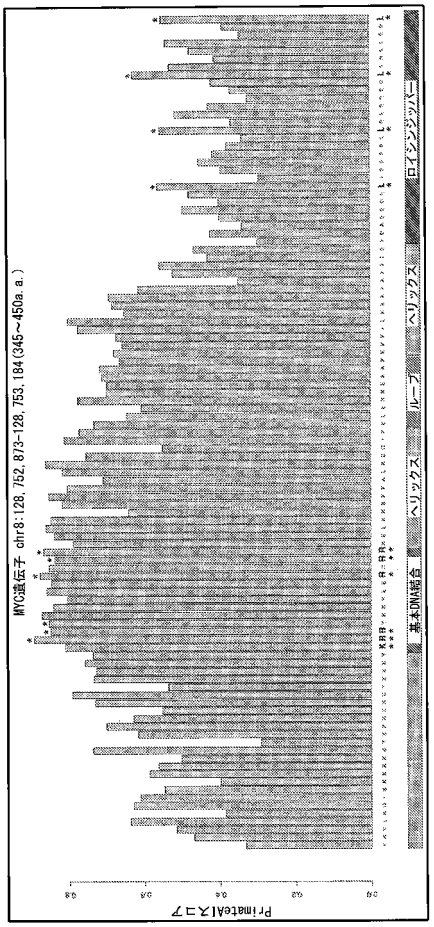
【図２５Ａ】



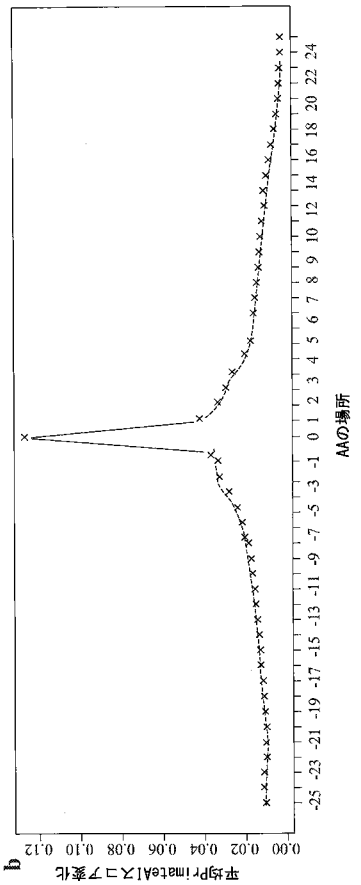
【図２５Ｂ】



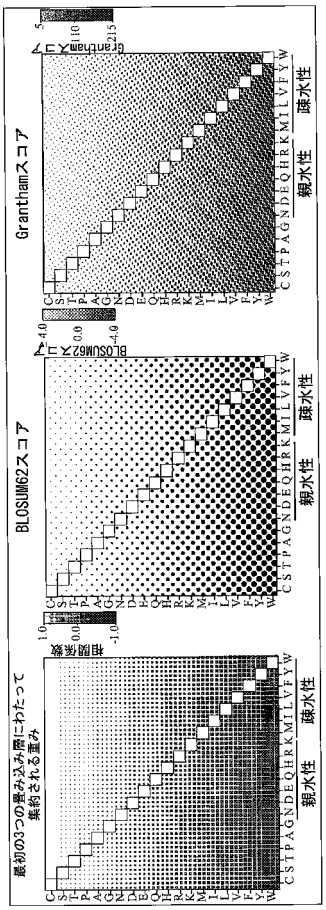
【図２５Ｃ】



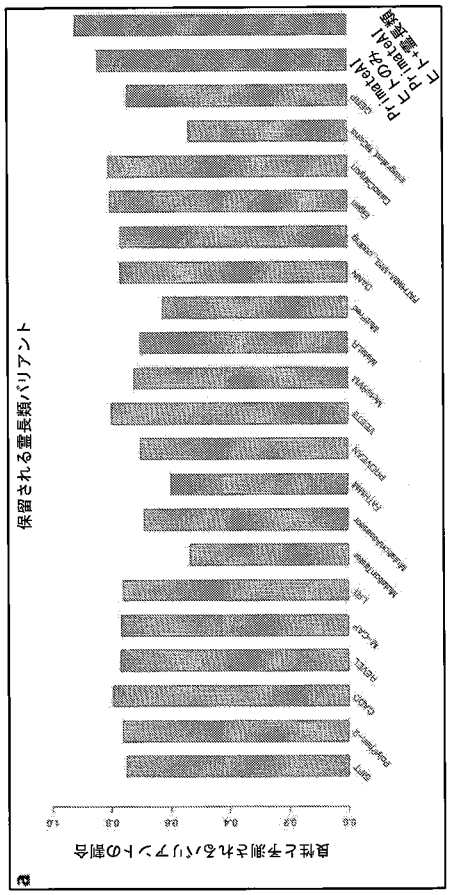
【図２６】



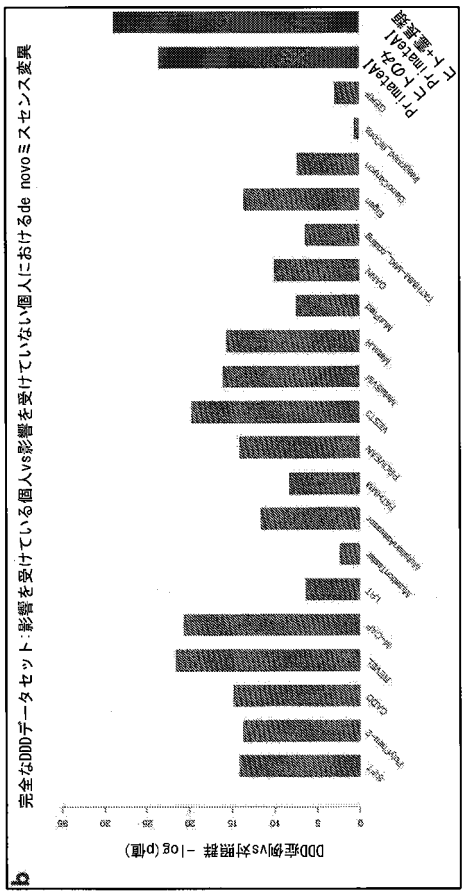
【図２７】



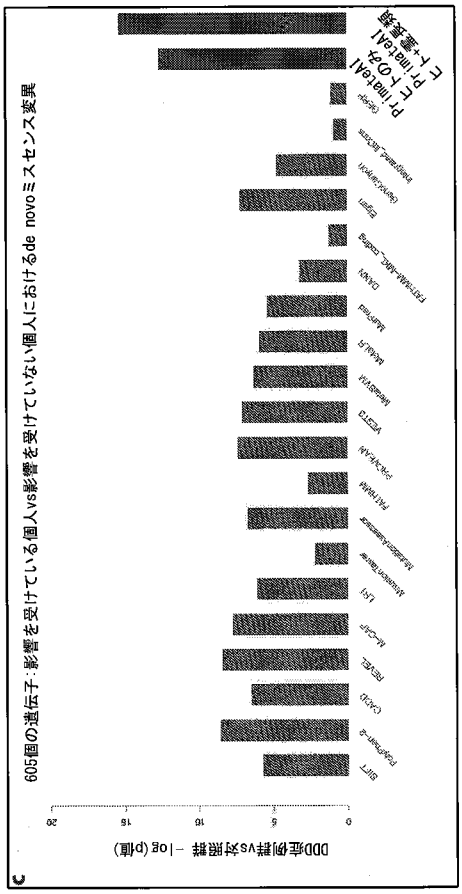
【図２８Ａ】



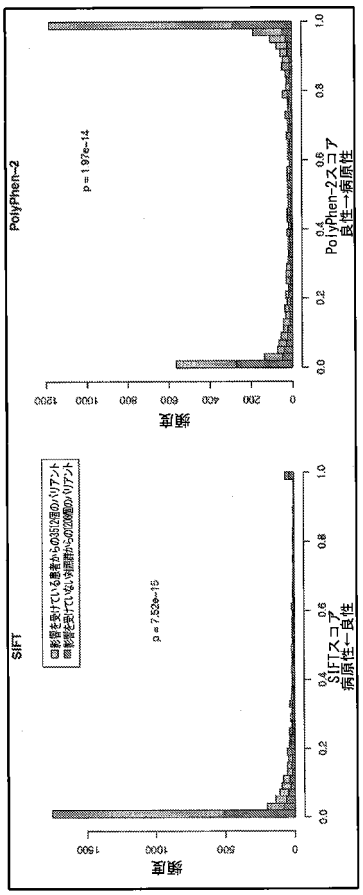
【図２８Ｂ】



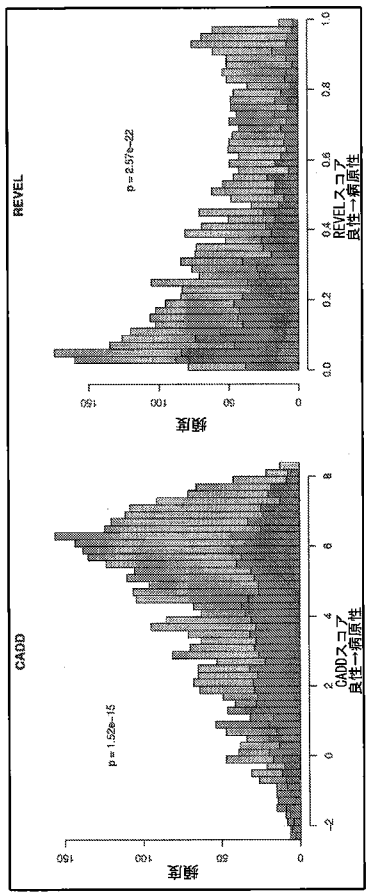
【図２８Ｃ】



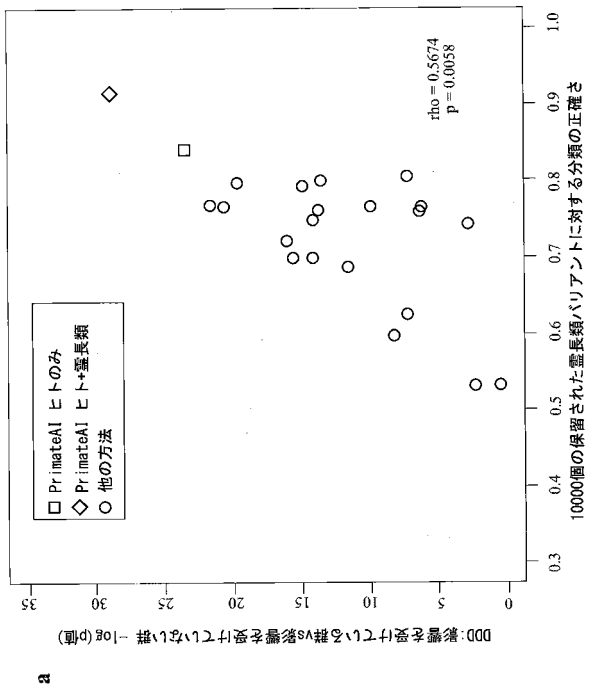
【図２９Ａ】



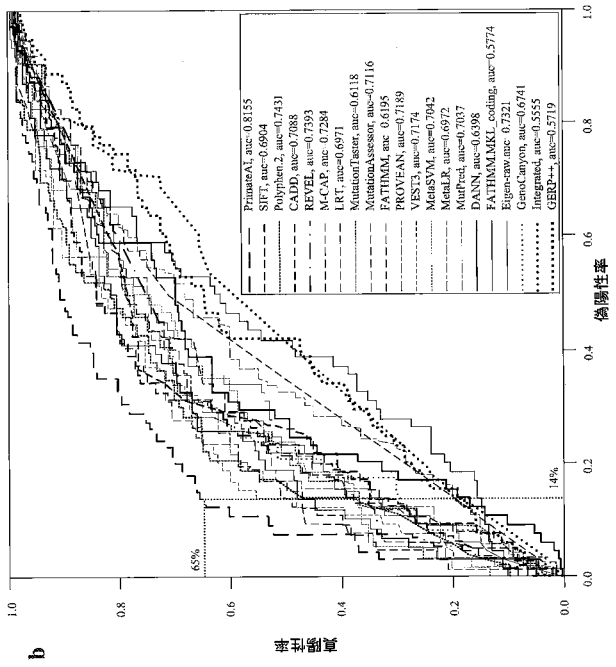
【図２９Ｂ】



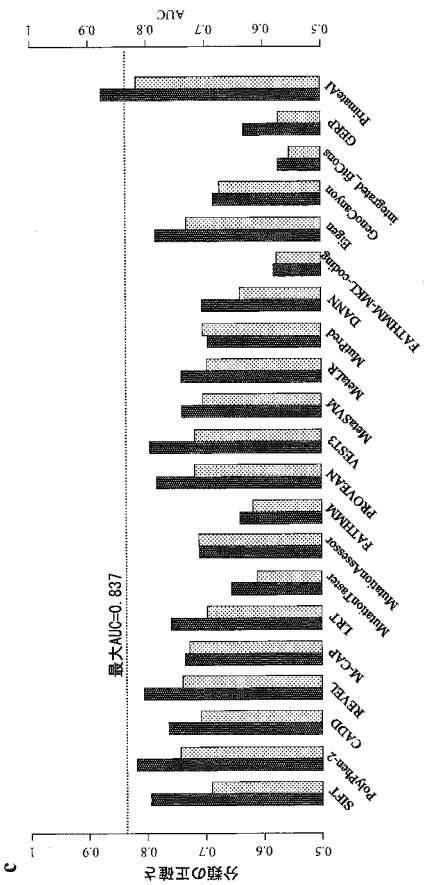
【図３０Ａ】



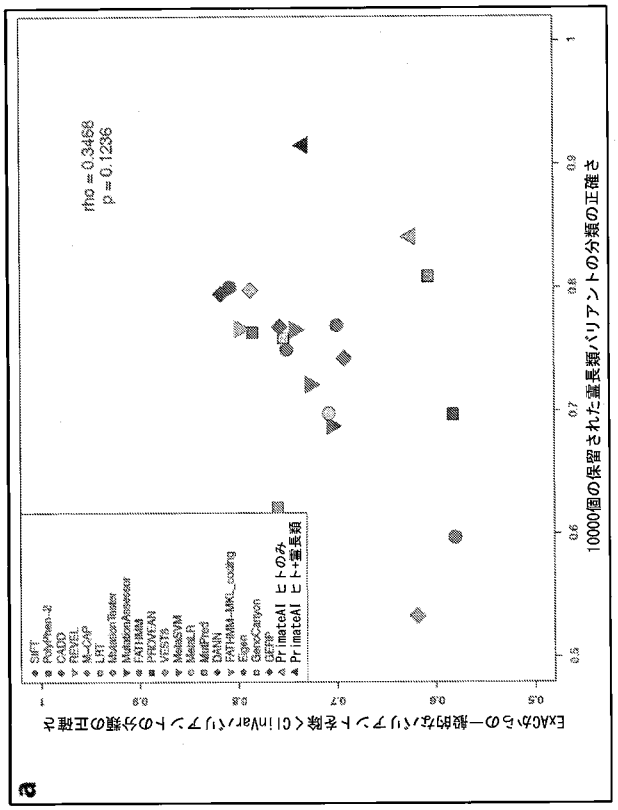
【図３０Ｂ】



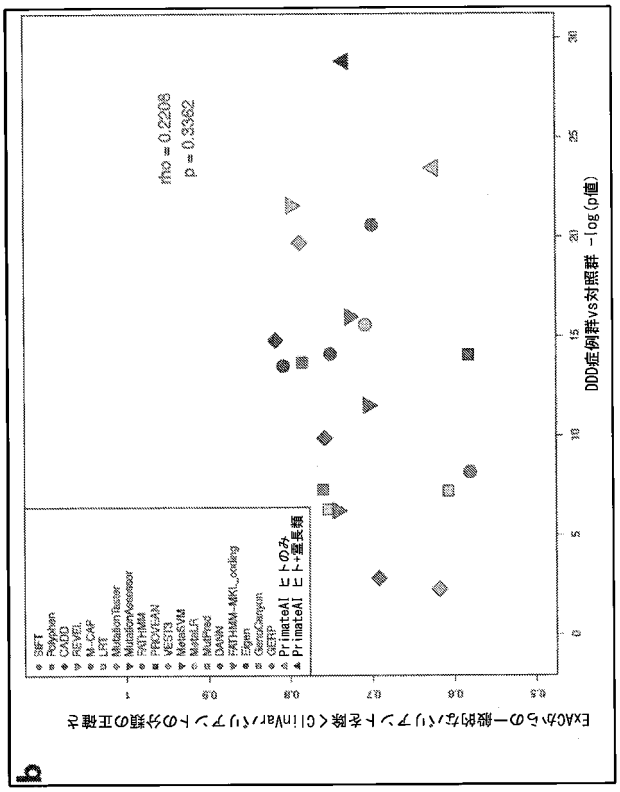
【図３０Ｃ】



【図３１Ａ】



【図３１Ｂ】



【図３２】

	訓練の正確さ	妥当性確認の正確さ	検定の正確さ
3状態二次構造	80.32 %	79.17 %	79.86 %
3状態溶媒接触性	64.83 %	60.53 %	60.31 %

【図３３】

モデル	反復	保留された霊長類バリアントの正確さ	DDD症例群vs対照群 −log (p値)	DDD 605個の疾患遺伝子 −log (p値)
DL予測される二次構造ラベルを使用する	1	0.914	28.60	15.61
	2	0.919	32.32	15.94
	3	0.915	29.48	16.38
	4	0.917	30.35	15.62
	5	0.916	29.11	16.22
	中央値	0.916	29.48	15.94
可能なときにDL予測される二次構造ラベルをヒトタンパク質のDSSP二次構造ラベルで置き換える	1	0.916	29.48	15.94
	2	0.913	31.72	16.24
	3	0.915	31.09	15.69
	4	0.915	30.49	16.09
	5	0.915	30.57	14.79
	中央値	0.915	30.79	16.09

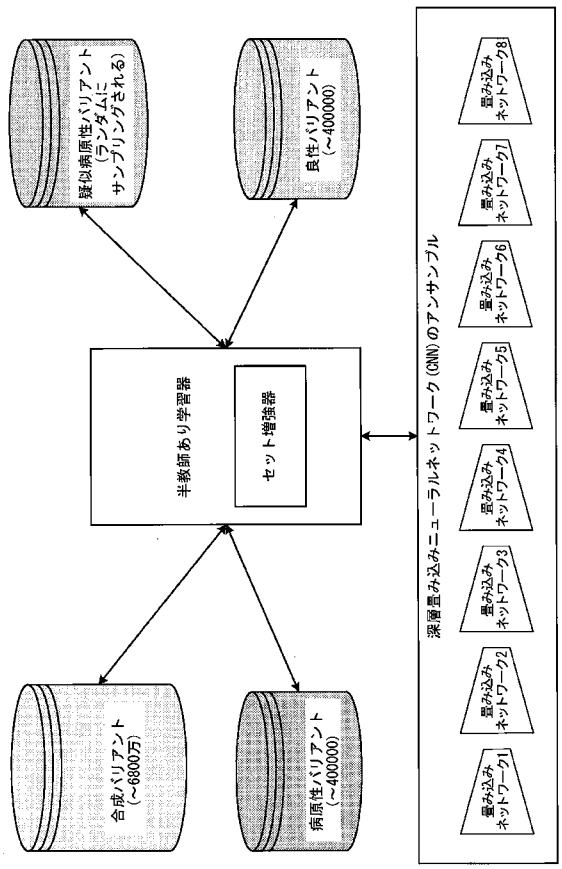
【図３４】

	保留された霊長類バリアントの正確さ	DDD症例群vs対照群 −log(p値)
SIFT	0.7490	14.1233
PolyPhen-2	0.7618	13.7044
CADD	0.7939	14.8167
M-CAP	0.7667	20.6078
LRT	0.7598	6.2960
MutationTaster	0.5332	2.2811
MutationAssessor	0.6879	11.5202
FATHMM	0.5981	8.1690
PROVEAN	0.6995	14.0747
VEST3	0.7978	19.7068
MetaSVM	0.7215	15.9829
MetaLR	0.6991	15.5842
REVEL	0.7686	21.5892
MutPred	0.6240	7.3044
DANN	0.7666	9.8976
MKL_coding	0.7657	6.2469
Eigen	0.8003	13.5364
GenoCanyon	0.8058	7.2196
integrated_fitCons	0.5331	0.4954
GERP	0.7430	2.8285
PrimateAI ヒト	0.8411	23.4536
PrimateAI ヒト+霊長類	0.9156	28.8346

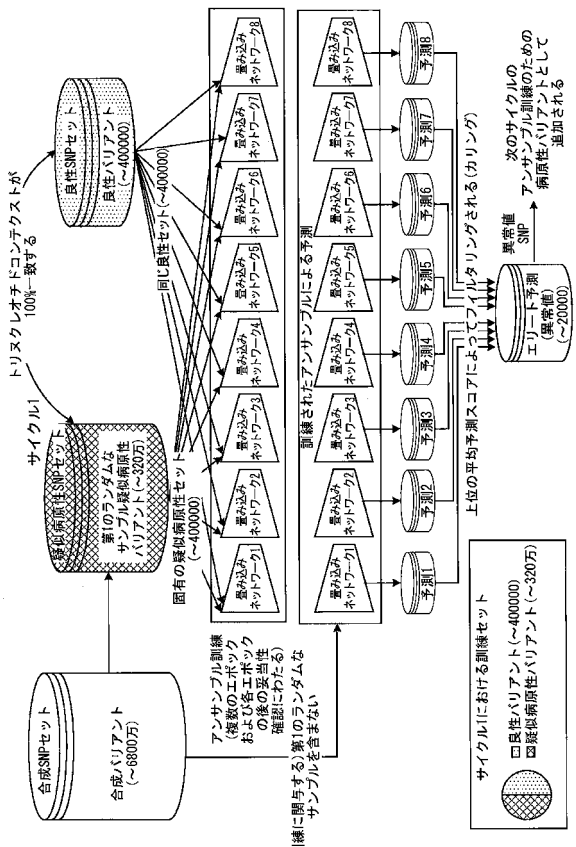
【図３５】

	−log (p値)	AUC	閾値	分類の正確さ
SIFT	5.6793	0.6903	0.001	0.7949
PolyPhen-2	8.5522	0.7430	0.956	0.8197
CADD	6.4772	0.7087	5.040	0.7643
M-CAP	7.7143	0.7283	0.081	0.7348
LRT	6.0285	0.6970	1.00E-06	0.7604
MutationTaster	2.1366	0.6117	1.000	0.6564
MutationAssessor	6.6779	0.7115	2.240	0.7096
FATHMM	2.5979	0.6194	-1.260	0.6401
PROVEAN	7.3447	0.7188	-4.310	0.7833
VEST3	7.0357	0.7173	0.679	0.7953
MetaSVM	6.2556	0.7041	-0.3371	0.7406
MetaLR	5.8707	0.6971	0.349	0.7407
REVEL	8.4115	0.7392	0.456	0.8056
MutPred	5.3370	0.7036	0.526	0.6949
DANN	3.1588	0.6397	0.996	0.7046
MKL_coding	1.1649	0.5773	0.965	0.5840
Eigen	7.1659	0.7320	0.577	0.7844
GenoCanyon	4.6763	0.6740	0.999	0.6875
integrated_fitCons	0.8315	0.5554	0.706	0.5736
GERP	1.0330	0.5718	5.060	0.6330
PrimateAI ヒト+霊長類	15.3064	0.8154	0.802	0.8772

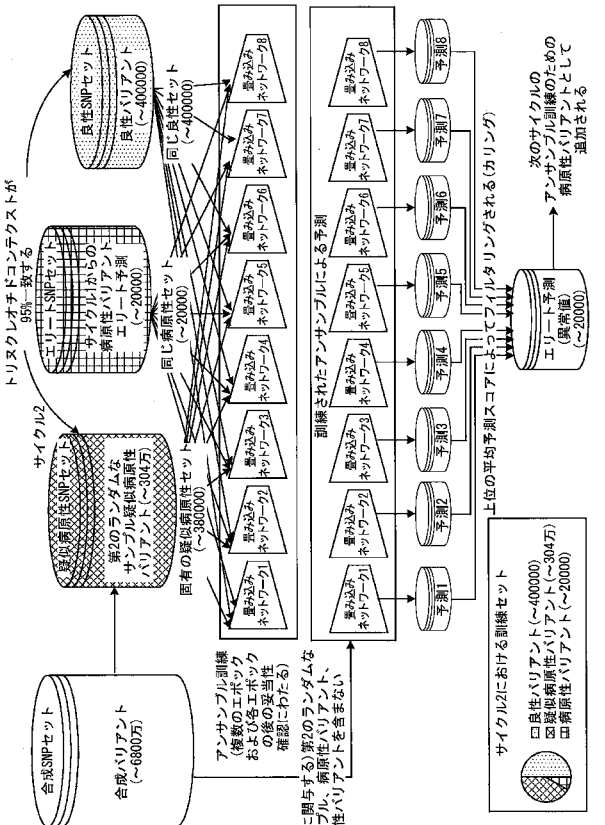
【図３６】



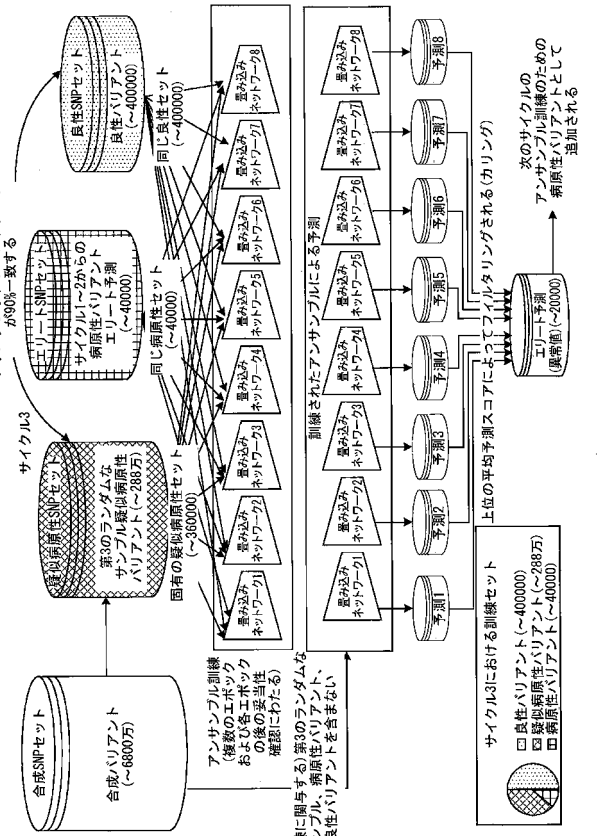
【図３７】



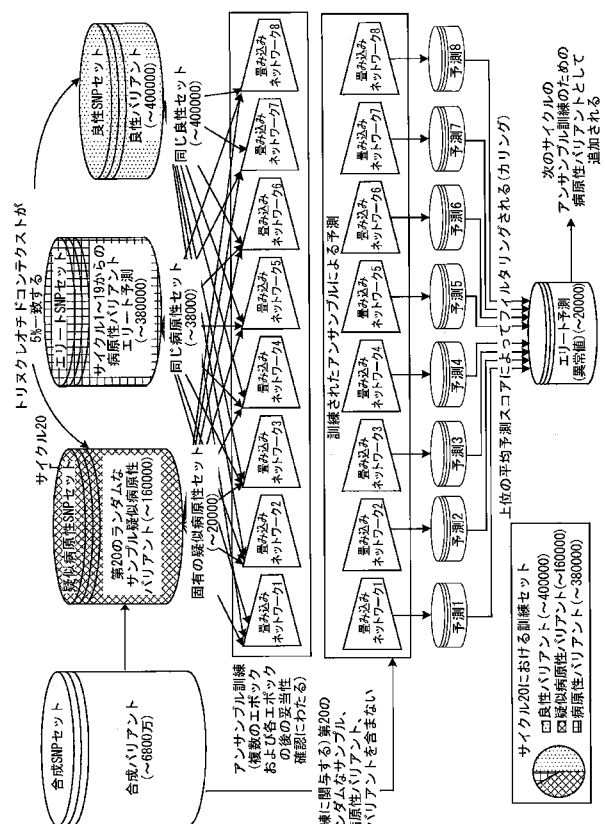
【図３８】



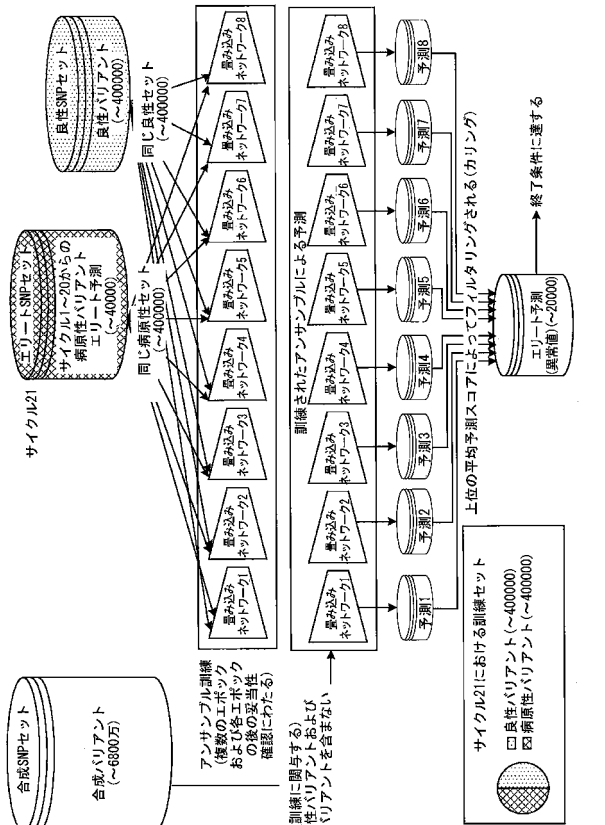
【図３９】



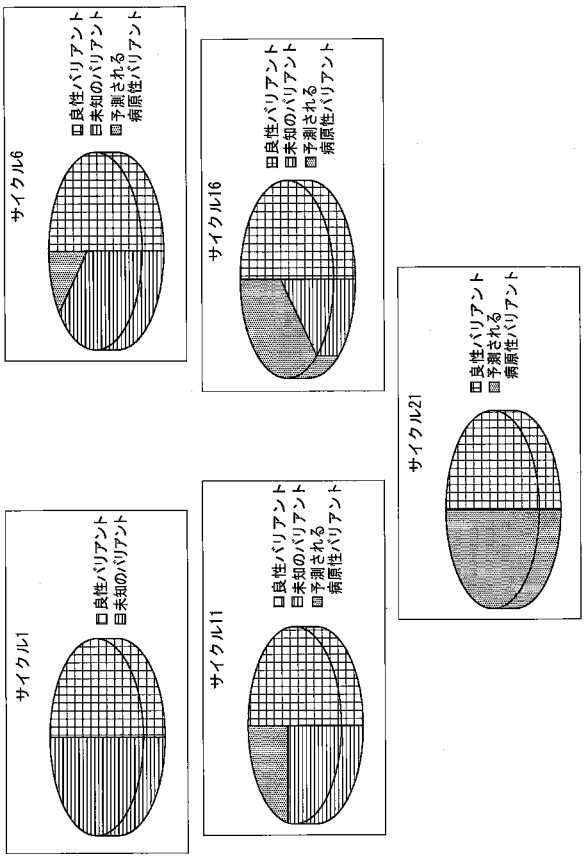
【図４０】



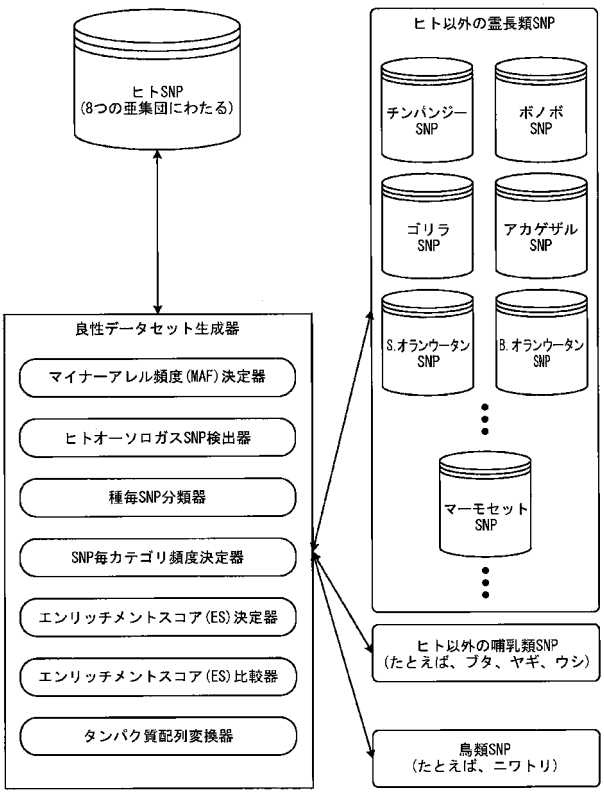
【図４１】



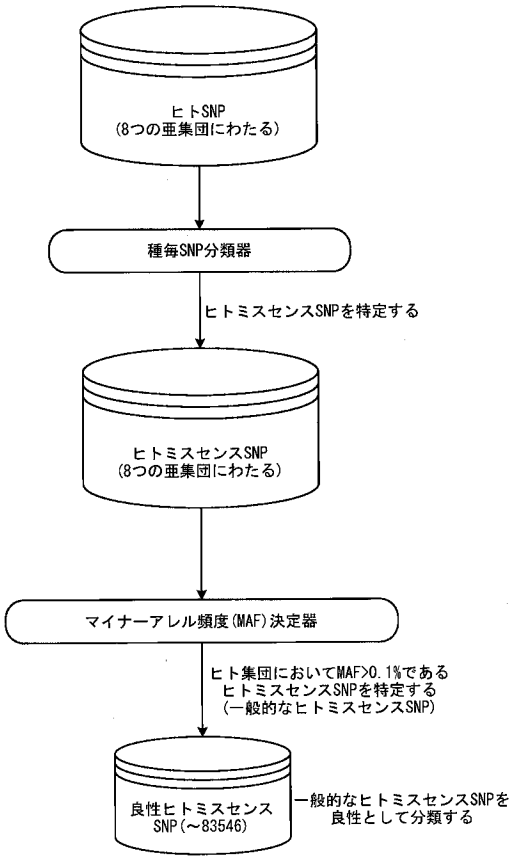
【図４２】



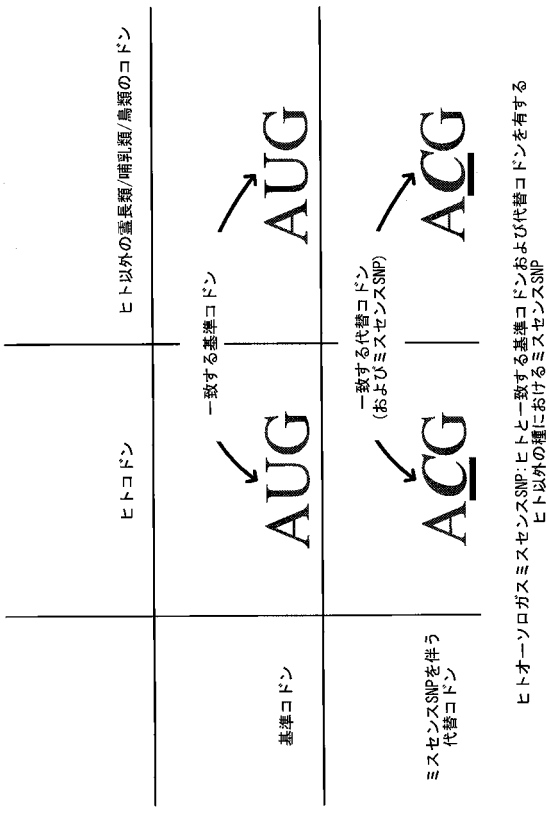
【図４３】



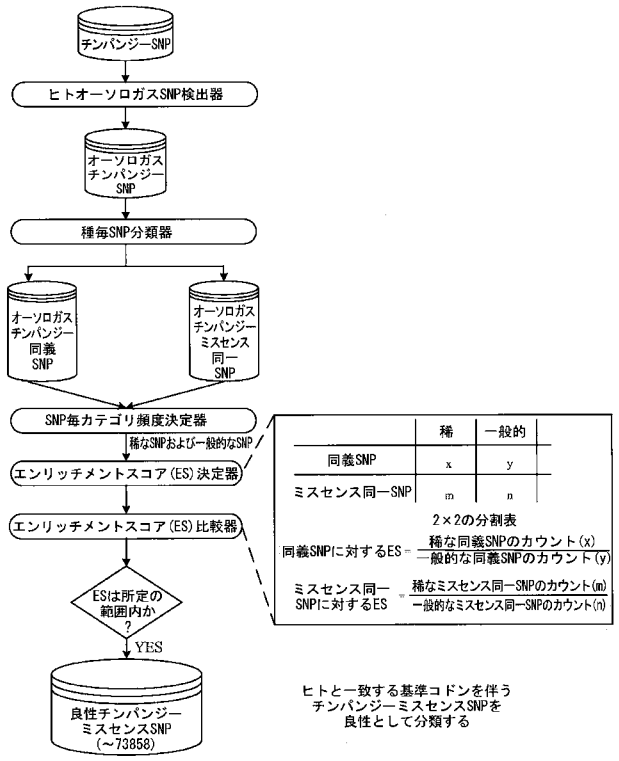
【図４４】



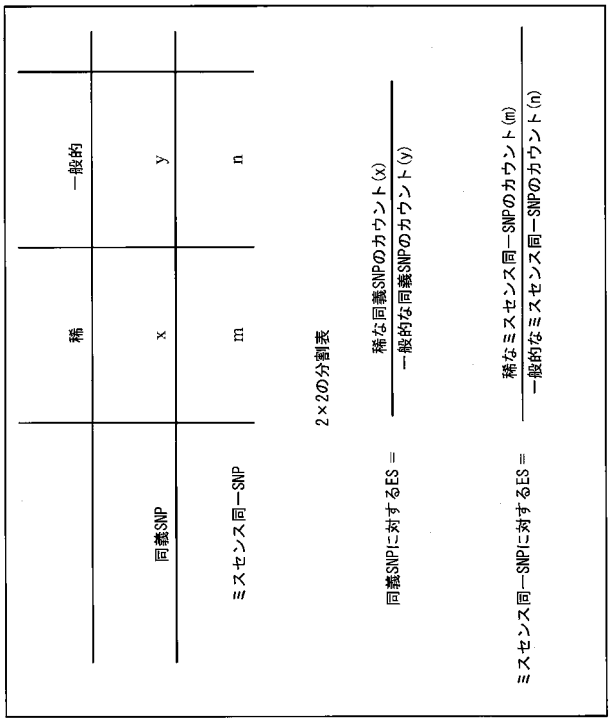
【図４５】



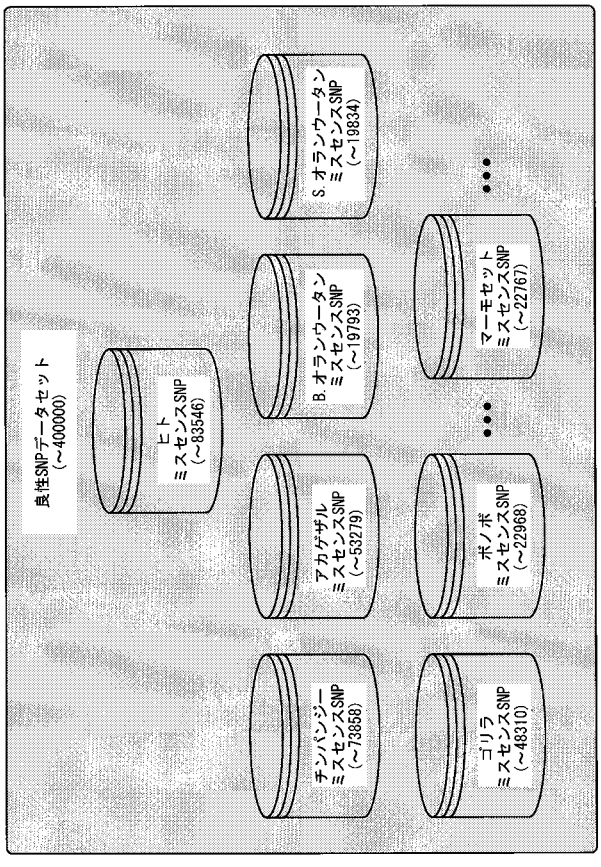
【図４６】



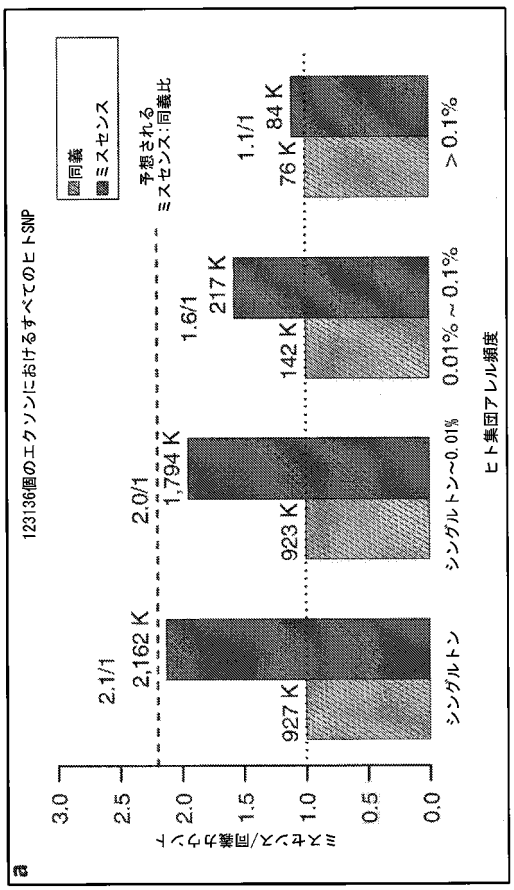
【図４７】



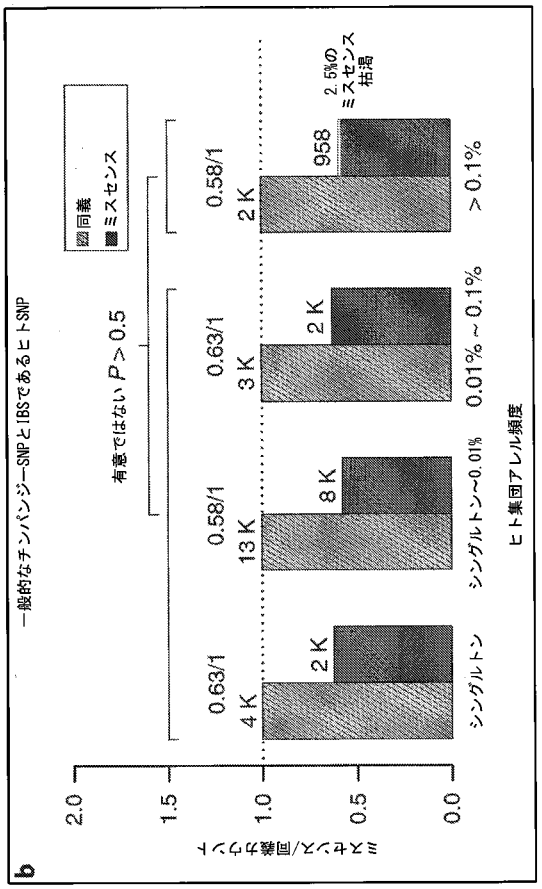
【図４８】



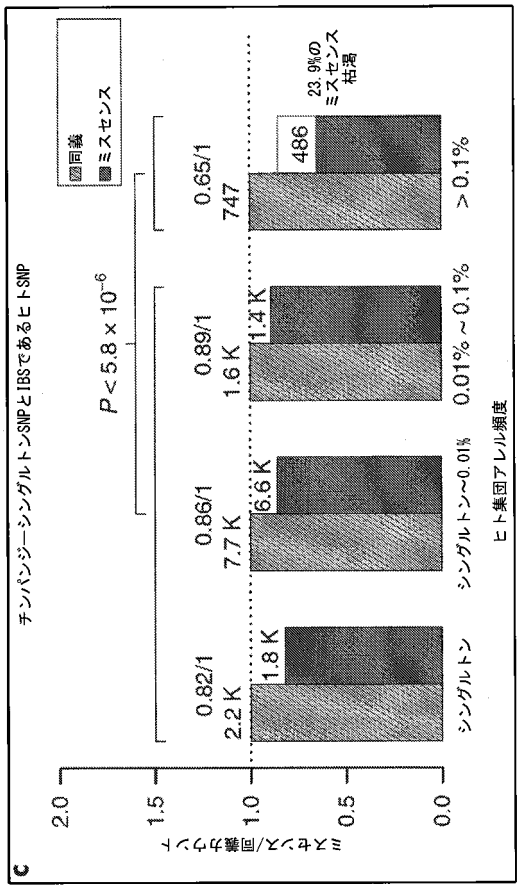
【図４９Ａ】



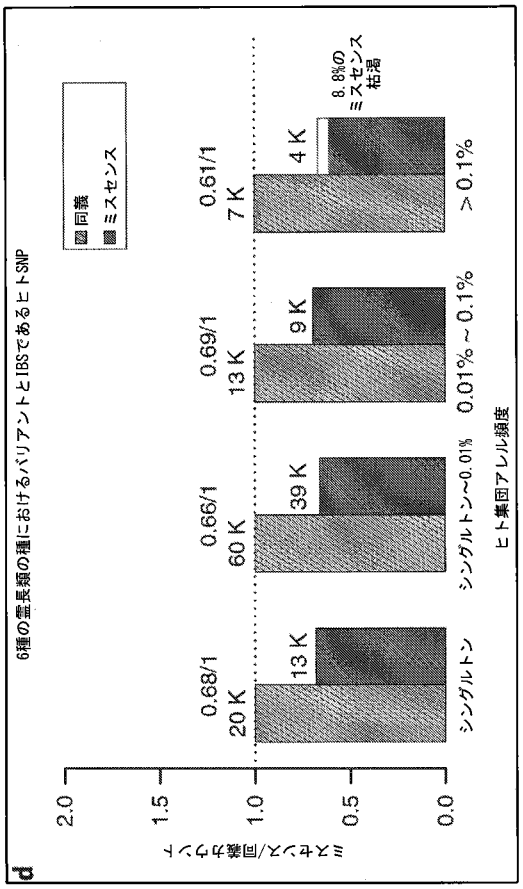
【図４９Ｂ】



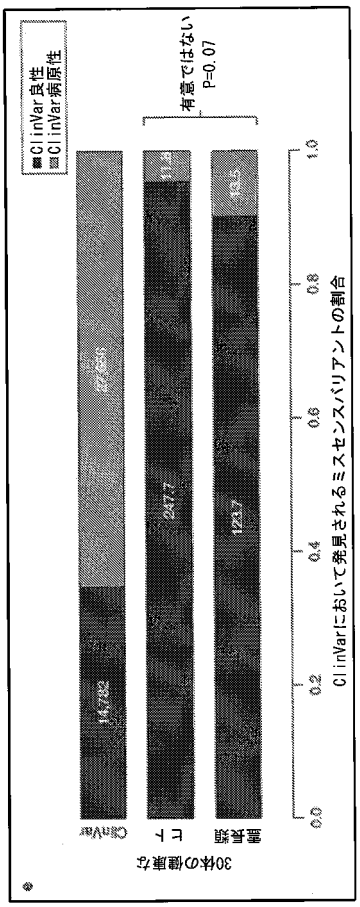
【図４９Ｃ】



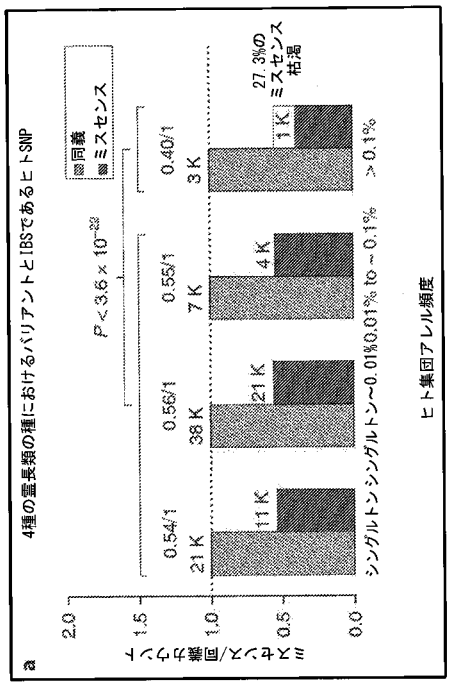
【図４９Ｄ】



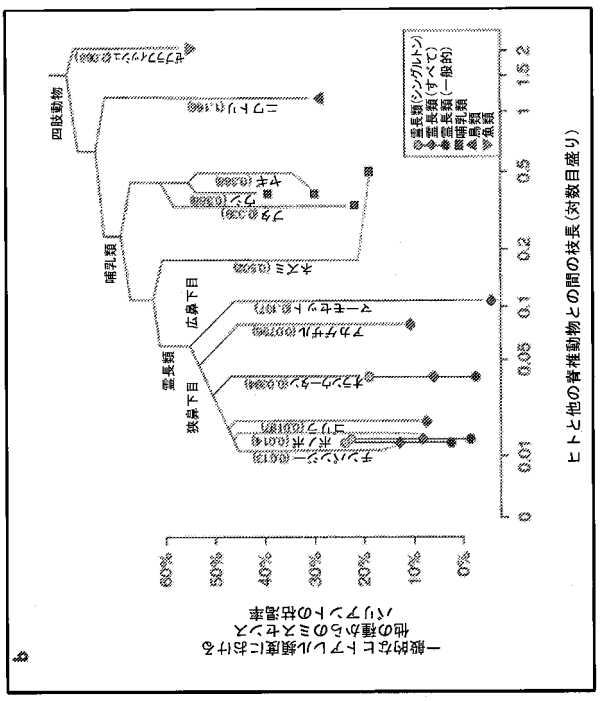
【図４９Ｅ】



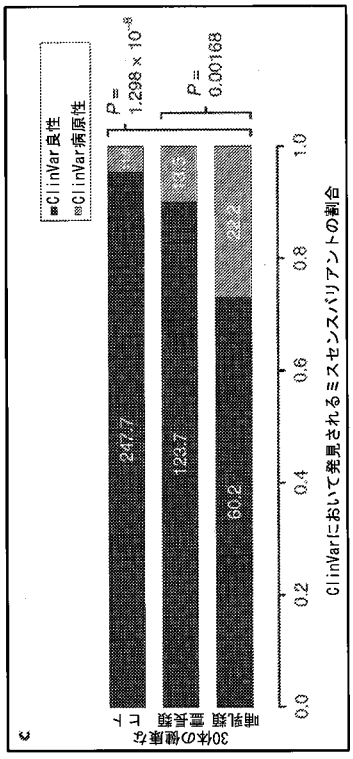
【図５０Ａ】



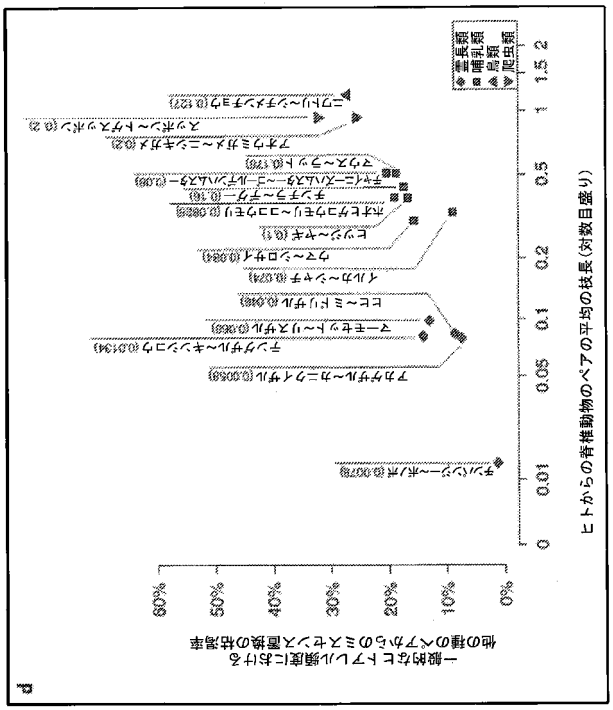
【図５０Ｂ】



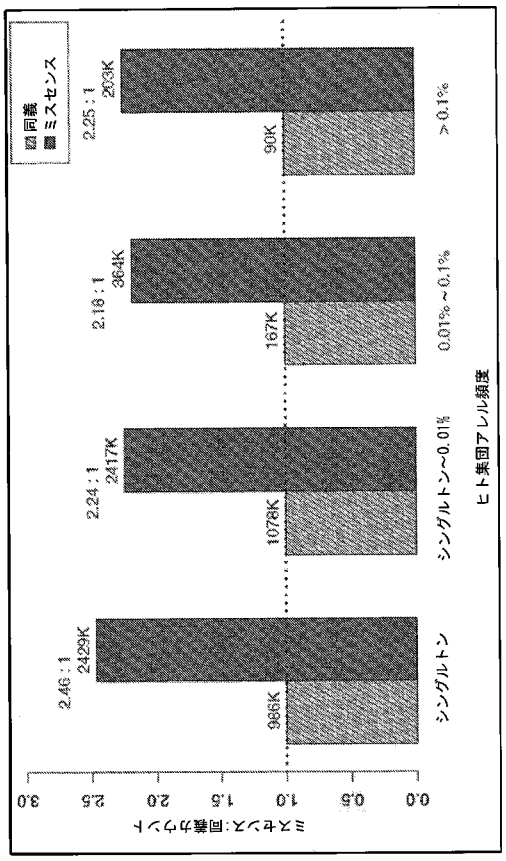
【図５０Ｃ】



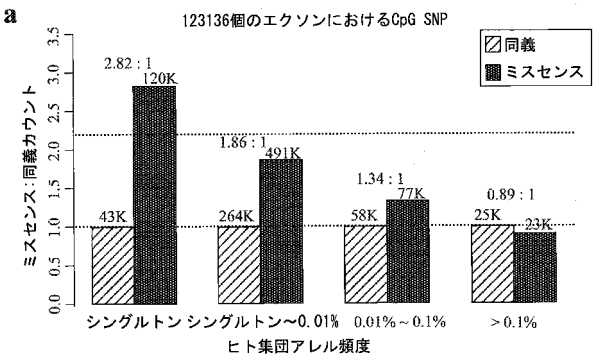
【図５０Ｄ】



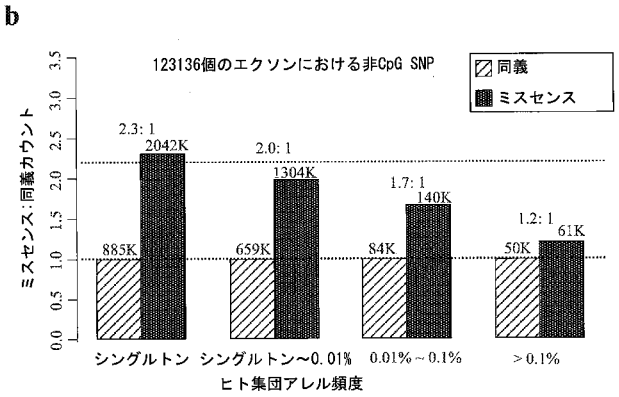
【図５１】



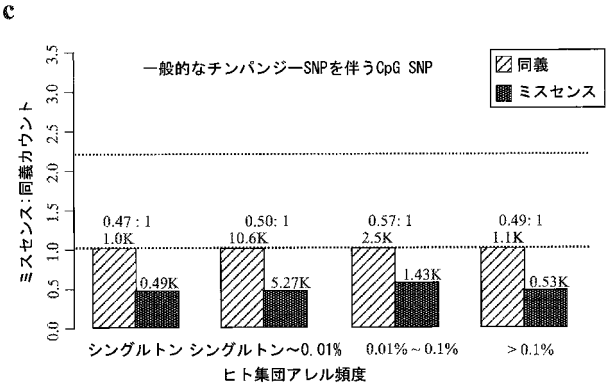
【図５２Ａ】



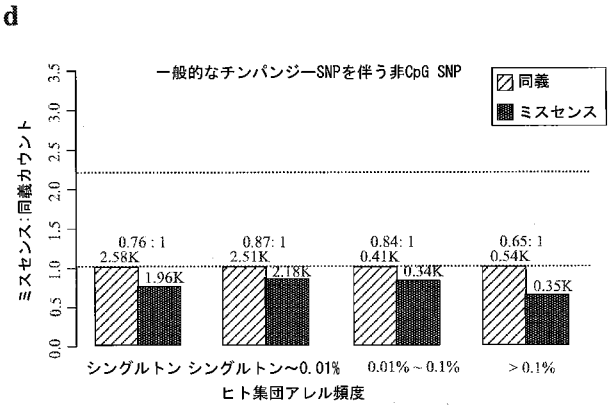
【図５２Ｂ】



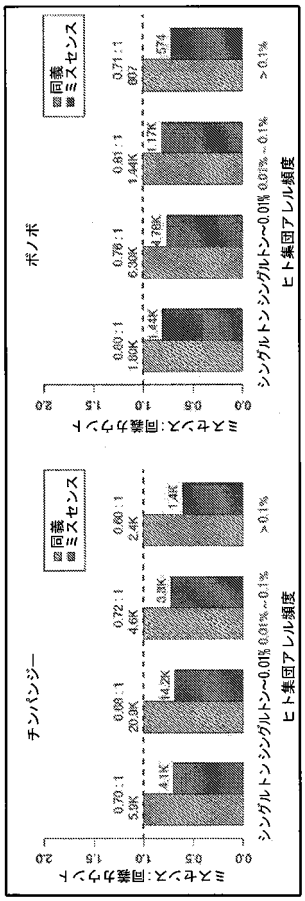
【図５２Ｃ】



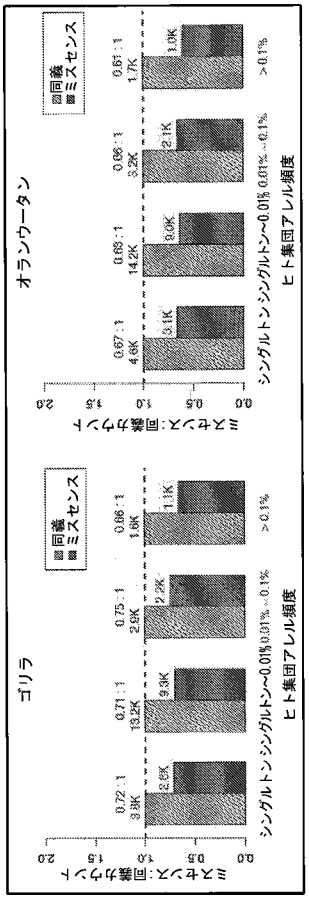
【図５２Ｄ】



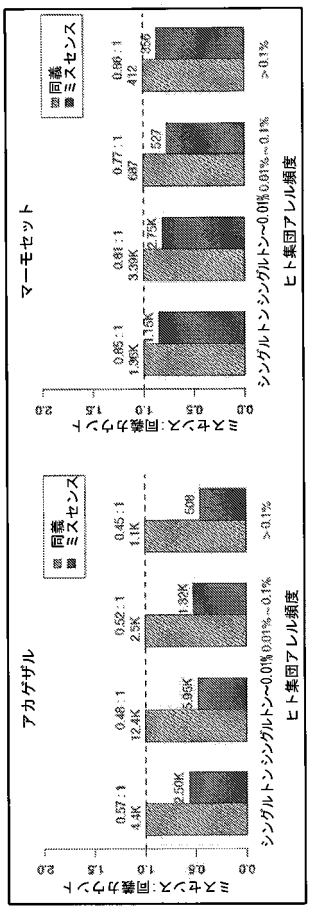
【図５３】



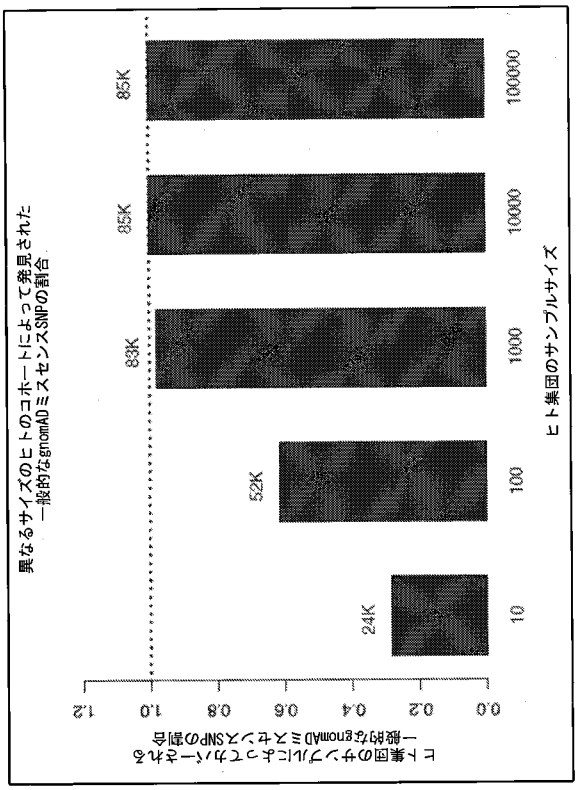
【図５４】



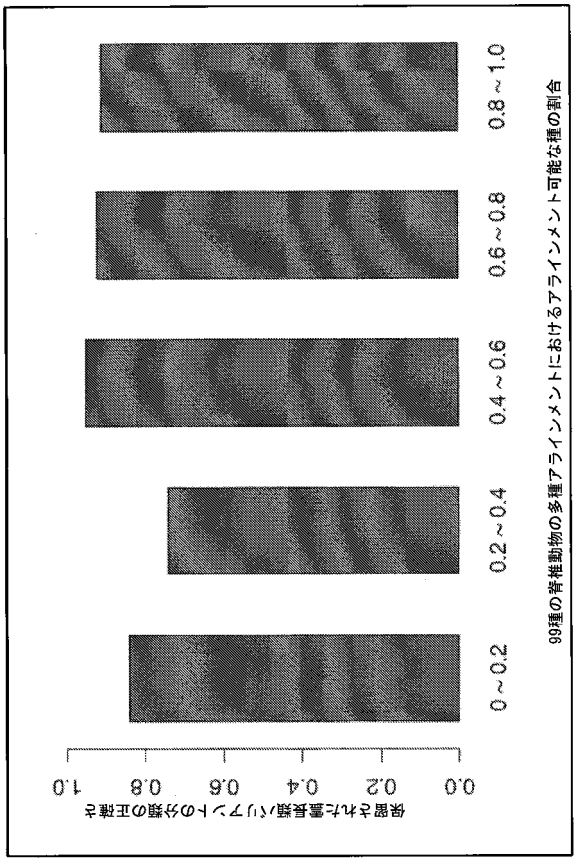
【図５５】



【図５６】



【図５７】



【図５８】

種 （および個体の数）	源	ヒトコドンと 一致するミスセンス バリアントの数	寄与する 固有のミスセンス バリアントの数
一般的な ヒトバリアント （MAF>0.1%、 123136人の個人）	gnomAD/ExAC データベース	83,546	83,546
チンパンジー（24）	Prado-Martinez他、 Nature（2013年）	76,293	73,858
ボノボ（13）	Prado-Martinez他、 Nature（2013年）	25,277	22,968
ゴリラ（27）	Prado-Martinez他、 Nature（2013年）	52,052	48,310
オランウータン、 ボルネオ亜種（5）	Prado-Martinez他、 Nature（2013年）	21,621	19,793
オランウータン、 スマトラ亜種（5）	Prado-Martinez他、 Nature（2013年）	27,567	19,834
チンパンジー（35）	de Manuel他、 Science（2016年）	75,580	30,327
オランウータン （個体データ 利用不可能）	dbSNP	18,627	10,554
アカゲザル （16、個体データ 利用不可能）	dbSNP	62,393	53,279
マーモセット （9、個体データ 利用不可能）	dbSNP	25,048	22,767
合計			385,236

【図５９】

123136個の エクソンにおける ヒト集団アレル頻度	同義バリアントの 予想される数	ミスセンスバリアントの 予想される数	ミスセンス：同義比
シングルトン	986,141.22	2,429,287.85	2.46
シングルトン~0.01%	1,077,630.37	2,416,751.61	2.24
0.01%~0.1%	166,849.26	364,292.75	2.18
>0.1%	90,221.42	202,918.38	2.24

【図６０】

	対象のClinVar バリアントの総数
元のClinVarファイル	666,403
hg19記録のみを保持する	324,698
–ヌクレオチドバリアントのみを抽出する	264,623
非コーディングバリアントを除去する	186,769
すべての同義変化と終止コドンを作成または破壊するミスセンス変化とを無視する	122,884
有意性が知られていないバリアントおよび矛盾するアノテーションを伴うバリアントを除外する	42,438

【図６１】

種	ClinVarにおける良性アノテーションを伴うバリアントの数	ClinVarにおける病原性アノテーションを伴うバリアントの数
チンパンジー	218	21
ボノボ	85	7
ゴリラ	167	23
オランウータン	160	12
アカゲザル	87	11
マーモセット	25	7
ネズミ	30	4
ブタ	74	30
ウシ	77	39
ヤギ	60	16
ニワトリ	9	8
ゼブラフィッシュ	2	6

【図６２】

表1|PrimateAIスコアが0.803以上であるミスセンスde novo変異(DNM)だけを考慮すると者に知的障害においてゲノムワイド有意性を達成する追加の遺伝子

HGNCシンボル	タンパク質切断バリアント	ミスセンス	PrimateAIスコア≧0.803	すべてのミスセンス	PrimateAIスコア≧0.803	すべてのミスセンス	複数の個人において観察される表現型の異常
ACTL6B	0	3	1	5	1.5×10^{-7}	2.4×10^{-6}	小頭症
EBF3	3	3	5	2	5.2×10^{-6}	5.4×10^{-6}	生育遅延、眼奇形、斜視、運動失調
EFTUD2	2	4	1	5	1.5×10^{-6}	1.5×10^{-5}	小頭症、耳介低位、小耳症、後鼻孔閉鎖
HECW2	1	8	2	8	2.8×10^{-6}	6.7×10^{-5}	発作、ミオパチー、頭蓋冠の異常
KDM6A	2	3	2	3	2.3×10^{-6}	9.8×10^{-6}	まぶた、歯牙の異常、筋緊張低下
KESC	0	3	3	3	3.0×10^{-6}	2.5×10^{-5}	知的発育不全
MAP2K7	0	5	3	5	3.1×10^{-6}	2.7×10^{-5}	両眼隔離、耳介低位、羊水過多
PPP2CB	0	6	1	6	1.5×10^{-6}	1.6×10^{-5}	顔の異常、低身長
PRKD1	0	6	2	6	8.6×10^{-6}	1.2×10^{-5}	肌、指、および心臓の異常、薄毛
SOX11	1	3	1	3	3.1×10^{-6}	2.4×10^{-5}	遠視、爪発育不全
TBR1	4	4	3	4	1.3×10^{-5}	4.2×10^{-5}	自閉症的行動
TLK2	3	5	1	5	4.7×10^{-6}	6.3×10^{-5}	鼻、まぶたの異常、傾斜した眼瞼裂
TRIM2	5	2	4	4	1.4×10^{-5}	5.4×10^{-5}	関節弛緩
UZAF2	0	4	1	4	2.6×10^{-6}	1.2×10^{-5}	発作、眼、口蓋、人中の異常

タンパク質切断DNMおよびミスセンスDNMのカウントが提供される。PrimateAIスコアが0.803以上であるミスセンス変異だけを用いて統計的検定が行われたとき、およびそれがすべてのミスセンス変異について繰り返されたときの、遺伝子エンリッチメントに対するP値が示される。

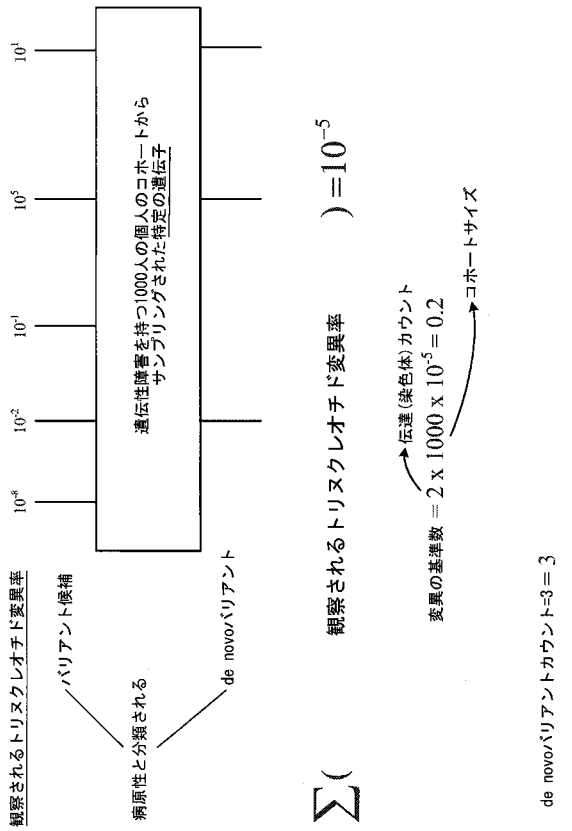
【図６３】

表2|ClinVarにおける専門家によりアノテートされたバリアントと、DDD症例群vs対照群におけるde novoバリアントとの間での、Granthamスコア、タンパク質表面露出率、およびアミノ酸配列保存率の差の比較

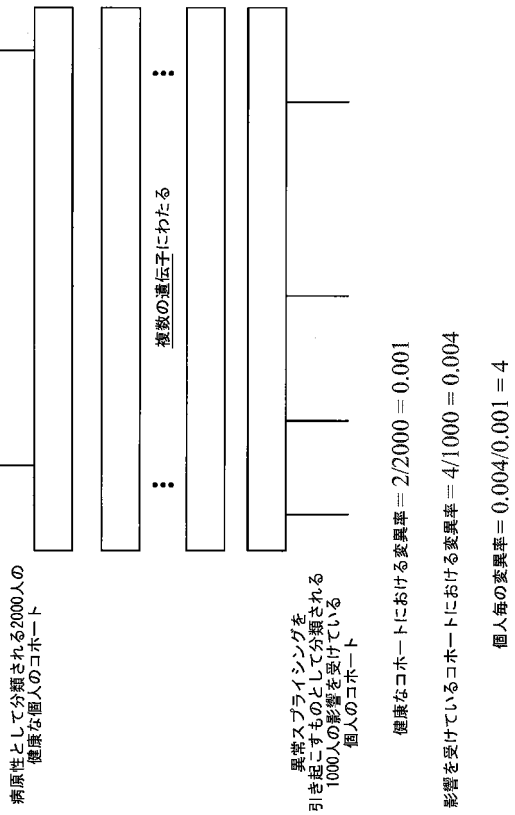
	Grantham スコア	タンパク質表面露出率	配列保存率
ClinVar病原性バリアント	91.1	0.53	0.87
ClinVar良性バリアント	67.4	0.41	0.54
専門家のアノテーションにおける差	+23.7	+0.12	+0.33
DDD患者におけるde novoバリアント	84.9	0.51	0.90
健康な対照群におけるde novoバリアント	72.7	0.29	0.73
影響を受けている個人と影響を受けていない個人の差	+12.2	+0.22	+0.17

ClinVarデータベースの中の矛盾しないアノテーションを伴うミスセンス変異、および605個の疾病関連遺伝子内でDDD症例群vs対照群において存在するde novoバリアントに対する、平均スコアが示される。タンパク質表面露出率は、溶媒接触性ニューラルネットワークによって露出している残基として予測されるアミノ酸の割合を反映し、配列保存率は、100種の脊椎動物のアラインメントにおいて配列相同性を有するアミノ酸の割合を示す。太字の数字は、経験的データと比較して、専門家により好まれている経験則における差を強調する。

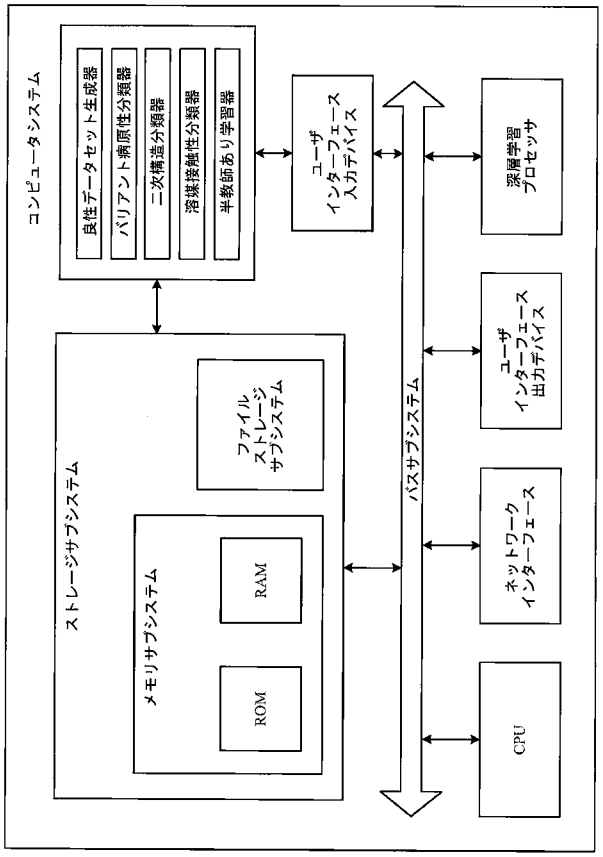
【図６４】


遺伝子毎のエンリッチメント分析
観察されるトリヌクレオチド変異率
バリアント候補
病原性と分類される
de novoバリアント
遺伝性障害を持つ1000人の個人のコホートから
サンプリングされた特定の遺伝子
観察されるトリヌクレオチド変異率
伝達(染色体)カウント
変異の基準数 = 2 x 1000 x 10^-5 = 0.2
コホートサイズ
de novoバリアントカウント=3 = 3

【図６５】


ゲノムワイドエンリッチメント分析
病原性として分類される2000人の
健康な個人のコホート
複数の遺伝子にわたる
異常スプライシングを
引き起こすものとして分類される
1000人の影響を受けている
個人のコホート
健康なコホートにおける変異率 = 2/2000 = 0.001
影響を受けているコホートにおける変異率 = 4/1000 = 0.004
個人毎の変異率 = 0.004/0.001 = 4

【図６６】


コンピュータシステム
良性データセット生成器
バリアント病原性分類器
二次構造分類器
溶媒接触性分類器
半教師あり学習器
ユーザ
インターフェース
入力デバイス
深層学習
プロセッサ
ストレージサブシステム
メモリサブシステム
ROM
RAM
ファイル
ストレージ
サブシステム
ユーザ
インターフェース
出力デバイス
バスサブシステム
ネットワーク
インターフェース
CPU

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2018/055840

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)
G06F

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, 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,P	SUNDARAM LAKSSHMAN ET AL: "Predicting the clinical impact of human mutation with deep neural networks", NATURE GENETICS, NATURE PUBLISHING GROUP, NEW YORK, US, vol. 50, no. 8, 23 July 2018 (2018-07-23), pages 1161-1170, XP036657698, ISSN: 1061-4036, DOI: 10.1038/S41588-018-0167-Z [retrieved on 2018-07-23] abstract page 1161 - page 1168 ----- -/--	1-23

[X] Further documents are listed in the continuation of Box C.

[X] 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	Date of mailing of the international search report
17 January 2019	25/01/2019

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Kürten, Ivayla

2

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2018/055840

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	QIONG WEI ET AL: "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics", PLOS ONE, vol. 8, no. 7, 9 July 2013 (2013-07-09), page e67863, XP055541955, DOI: 10.1371/journal.pone.0067863	1-5, 15-23
A	abstract page 1 - page 4	6-14
Y	CA 2 894 317 A1 (DEEP GENOMICS INCORPORATED [CA]) 15 December 2016 (2016-12-15)	1-5, 15-23
A	abstract paragraph [0032] - paragraph [0036] paragraph [0052] - paragraph [0057] paragraph [0119] - paragraph [0126]	6-14
A	IOANNIDIS NILAH M ET AL: "REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants", AMERICAN JOURNAL OF HUMAN GENETICS, AMERICAN SOCIETY OF HUMAN GENETICS, CHICAGO, IL, US, vol. 99, no. 4, 22 September 2016 (2016-09-22), pages 877-885, XP029761034, ISSN: 0002-9297, DOI: 10.1016/J.AJHG.2016.08.016 abstract page 1581 - page 1582	1-23
A	KARTHIK A JAGADEESH ET AL: "M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity", NATURE GENETICS., vol. 48, no. 12, 1 December 2016 (2016-12-01), pages 1581-1586, XP055541948, NEW YORK, US ISSN: 1061-4036, DOI: 10.1038/ng.3703 abstract page 878	1-23

2

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2018/055840

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
CA 2894317 A1	15-12-2016	NONE	

フロントページの続き

(31)優先権主張番号 62/573,153
(32)優先日 平成29年10月16日(2017.10.16)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号 62/582,898
(32)優先日 平成29年11月7日(2017.11.7)
(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,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,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,NE,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,JO,JP,KE,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,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT

(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・９２１２２・サン・ディエゴ・イルミナ・ウェイ・５２００
(72)発明者 ラクシュマン・サンダラム
アメリカ合衆国・カリフォルニア・９２１２２・サン・ディエゴ・イルミナ・ウェイ・５２００
(72)発明者 ジェレミー・フランシス・マクレー
アメリカ合衆国・カリフォルニア・９２１２２・サン・ディエゴ・イルミナ・ウェイ・５２００